



3 1761 10374377 9



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743779>

12
-001

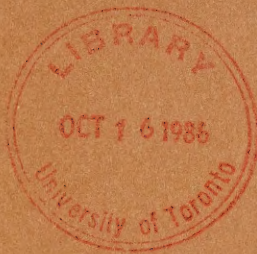


Statistics
Canada

Statistique
Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 15, NUMBER 1
JUNE 1989

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1989

Published under the authority of the Minister
of Regional Industrial Expansion

©Minister of Supply
and Services Canada 1989

Extracts from this publication may be reproduced
for individual use without permission provided the
source is fully acknowledged. However, reproduction
of this publication in whole or in part for purposes of
resale or redistribution requires written permission from
the Programs and Publishing Products Group, Acting
Permissions Officer, Crown Copyright Administration,
Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9

September 1989

Price: Canada, \$30.00 a year
Other Countries, \$35.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 15, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	M.N. Murthy, <i>Applied Statistics Centre, India</i>
L. Biggeri, <i>University of Florence</i>	W.M. Podehl, <i>Statistics Canada</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
W.A. Fuller, <i>Iowa State University</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.F. Gentleman, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
D. Holt, <i>University of Southampton</i>	V. Tremblay, <i>Statplus, Montreal</i>
	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

J. Armstrong, J. Gambino and J.-L. Tambay, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$30.00 per year in Canada, \$35.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US \$16.00 (\$20.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 15, Number 1, June 1989

CONTENTS

In This Issue	1
P.S. KOTT	
Robust Small Domain Estimation Using Random Effects Modeling	3
G.E. BATTESE, N.A. HASABELNABY and W.A. FULLER	
Estimation of Livestock Inventories Using Several Area and Multiple Frame Estimators	13
D.A. BINDER and J.P. DICK	
Modelling and Estimation for Repeated Surveys	29
J. BETHEL	
Sample Allocation in Multivariate Surveys	47
E.R. BRUNING and M.Y. HU	
The Role of Demographic Factors in the Analysis of Survey Versus Diary Purchase Reporting Accuracy	59
S. LEMESHOW and G. STROH JR.	
Quality Assurance Sampling for Evaluating Health Parameters in Developing Countries	71
Special Section – Statistical Uses of Administrative Data	
P. REDFERN	
European Experience of Using Administrative Data for Censuses of Population: The Policy Issues That Must be Addressed	83
W.E. WINKLER	
Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage	101
J.R. JONAS and P.S. HANCZARYK	
Automated Quality Assurance Processing of Administrative Record Files	119
J.C. MOORE and K.H. MARQUIS	
Using Administrative Record Data to Evaluate the Quality of Survey Estimates ...	129
C. CLARK and R. LUSSIER	
The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities	145

In This Issue

This issue of **Survey Methodology** contains a special section on the **Statistical Uses of Administrative Data**. The five papers in the section cover a diversity of topics ranging from policy issues to data processing.

With the increasing emphasis on the use of administrative records by statistical agencies, probabilistic matching or record linkage methods are becoming more widespread. Most applied work is done using the framework described by Fellegi and Sunter (1969). Winkler examines the importance of an independence assumption that is usually employed in applications involving the Fellegi-Sunter model because it leads to great computational simplification. In the context of a problem involving matching lists of businesses he investigates modifications that can be used when the independence assumption is not valid.

The paper by Redfern deals with a statistical use of administrative records that is of great importance for statistical agencies — the use of administrative records as a source of census data. He notes that Denmark has completely abandoned the traditional questionnaire-based census in favour of the use of administrative records to obtain census data. In three other European countries, some data that were traditionally collected using a census questionnaire are now obtained directly from administrative sources. The author considers the situation in the United Kingdom in detail. He concludes that public concerns about invasion of privacy, as well as political ideology and scarce resources, are blocking the consolidation of administrative information from a number of diverse sources into a central population register. He suggests that, although political considerations will always carry the greatest weight in any discussion of the development of a population register, statisticians have an obligation to make their views known.

Jonas and Hanczaryk note that the role of administrative data at the U.S. Bureau of the Census has increased over time. The need for an overall quality management system that is responsive to problems related to the processing of very large amounts of data was recognized before the 1987 Economic Censuses. The system that was developed involves the extensive use of microcomputers to reduce costs.

Moore and Marquis describe an application involving the use of administrative data in survey evaluation. Information from the Survey of Income and Program Participation conducted by the U.S. Bureau of the Census was matched to administrative records for five federal programs and four state programs using record linkage methods. Analysis of the data set is just beginning. The objectives of the study are to quantify the effects of measurement errors and to use this information to derive more efficient survey designs.

Statistics Canada is in the process of reorganizing its programme of economic surveys. A key element is the rebuilding of its central register of economic entities, which will serve as the frame for economic surveys. Clark and Lussier's paper outlines the concepts and procedures underlying the establishment and maintenance of profiles of economic entities and describes the role of administrative data in this task. A number of issues with respect to profiling activities are raised following a simulation study.

In this issue's first paper, Kott develops a small domain estimator which meets the criterion of design consistency introduced by Isaki and Fuller (1982). The mean squared error of this estimator is evaluated. Using an empirical example, Kott shows that the mse estimator can be used to choose between the proposed small domain estimator and the conventional design-based estimator.

Published estimates for periodic surveys are often based only on the current sample, thereby failing to exploit correlations with estimates from previous periods. On the other hand, economists and other social scientists frequently ignore the sampling error when using these estimates in their time series models. Binder and Dick show how sampling error can be taken into account in these models. For readers new to the area, the authors provide a brief review of previous work, with an extensive list of references.

Battese, Hasabelnaby and Fuller investigate a procedure for constructing a composite estimator for livestock numbers. The authors use a linear model to pool six types of estimators from the U.S. Department of Agriculture June Enumerative Survey over several years. Empirical results show the improvements in variance for the optimal linear combination of the six estimators within a particular year, with further improvements if the other years' estimators are included.

Bethel examines optimal allocation for multipurpose surveys. A study of the sensitivity of the optimal allocation to changes in variance constraints is presented. Bethel derives results which can be used to determine if survey costs can be reduced significantly by allowing some variances to increase marginally. He also presents an iterative algorithm for solving the optimization problem.

Bruning and Hu provide insight into the issue of survey recall versus diary collection methods. They start with a literature review of studies and comparisons of the two methods. The main part of the paper deals with an experiment to assess the relationship between several demographic factors and the collection methods. The findings confirm those of earlier studies but also strongly raise the possibility of measurement problems with the survey recall collection method.

Quality assurance sampling is applied by Lemeshow and Stroh to the problem of reducing the sample size needed to ascertain whether a population meets certain health standards. The example used by the authors is the immunization coverage of children in developing countries. The sampling method uses an initial sample to test the hypothesis of adequate vaccination by stratum. Strata where the test result is not sufficiently conclusive are subjected to additional sampling.

The Editor

Robust Small Domain Estimation Using Random Effects Modeling

PHILLIP S. KOTT¹

ABSTRACT

This paper develops a design consistent small domain estimator using a random effects model. The mean squared error of this estimator is then evaluated *without* assuming the random effect component of the model is correct. Data from a complex sample survey shows how this approach to mean squared error estimation, while perhaps too instable to be used directly, can be employed to determine whether the design consistent small domain estimator proposed here is better than the conventional design-based estimator.

KEY WORDS: Finite population; Model; Mean squared error; Design consistent; Randomization.

1. INTRODUCTION

Suppose we were given a probability sample of unit values and were asked to estimate the mean of a small domain within the larger population covered by the sample. Scott and Smith (1969) introduced a Bayesian estimator for this purpose and showed that their estimator could also be developed using only unbiasedness and minimum variance (UMV) criteria. Their UMV approach, sometimes called random effects or components-of-variance modeling, will be adopted here.

Most attempts at small domain estimation paralleling Scott and Smith (*e.g.*, Fay and Herriot 1979, Battese and Fuller 1971, Ghosh and Meeden 1986, Prasad and Rao 1986, Fuller and Harter 1987, and Stroud 1987) assume that the sampling design is noninformative and so ignorable. The same assumption is made for synthetic estimators of small domain means, which will not be discussed at any depth here (for examples of these, see Gonzalez and Hora 1978).

Assuming a noninformative sampling design misses perhaps the most important contribution of randomization to inference. Since most statistical models in finite population inference are either wrong or (at best) incomplete, it is desirable for an estimation strategy to have the following property: if the sample were large enough, the estimator should approach what it is estimating almost certainly no matter what the "true" model. This desire receives formal expression in the criterion of design consistency introduced by Isaki and Fuller (1982).

Design consistency is an asymptotic property. As a result, it is often necessary to hypothesize a model (or models) when choosing among alternative design consistent estimation strategies. This is especially true in the case of small domain estimation, where the sample may be particularly small and the sampling design beyond one's control. Nevertheless, limiting attention to design consistent estimators does offer some, albeit small, protection against model failure. Using this reasoning, Särndal (1984) focused his attention on design consistent small domain estimators. We will follow that practice here.

Section 2 develops a design consistent random effects estimator for a small domain population mean. Section 3 introduces a robust (but unstable) estimator for the model and design mean squared errors of the small domain estimator. It is robust in the sense of not depending

¹ Phillip S. Kott, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, D.C., 20250, USA.

on the necessary, but heroic, model that links the small domains together. Section 4 contains an empirical example and Section 5 a discussion.

2. THE ESTIMATOR

We begin with the *basic* (or fixed effects) model:

$$y_{gi} = \theta_g + \epsilon_{gi}, \quad (1)$$

where the ϵ_{gi} are uncorrelated random variables with means of zero, and $\text{var}(\epsilon_{gi}) = \delta_g^2$. The subscript gi denotes a unit in domain g . There are N_g units in the population from domain g and m domains.

Let us focus on a particular domain j . The problem is to estimate the domain mean:

$$\bar{y}_{jP} = \sum_{i=1}^{N_j} y_{ji}/N_j.$$

Let p_{ji} be the probability of selecting unit ji for the sample and n_j be the number of units selected from domain j . It is well known that a design unbiased and model efficient linear estimation strategy for \bar{y}_{jP} would set the p_{ji} equal to n_j/N_j and the estimator equal to $\sum_{i=1}^{n_j} y_{ji}/n_j$, where the units are relabeled so that $j1, \dots, jn_j$ are in the sample.

Unfortunately, one is often required in practice to estimate a domain mean using a sample that has not been selected primarily for that purpose. Consequently, the selection probabilities within domain j may not all equal n_j/N_j . A popular estimator in this circumstance is

$$d_j = \sum_{i=1}^{n_j} w_{ji} y_{ji}, \quad (2)$$

where

$$w_{ji} = p_{ji}^{-1} / \sum_{k=1}^{n_j} p_{jk}^{-1},$$

denotes the *sampling weight* of unit ji . This estimator was suggested by Brewer (1963) and Hajek (1971).

The estimator d_j is clearly model unbiased under (1), in the sense that $E_\epsilon(d_j - \bar{y}_{jP}) = 0$. Under many sampling designs, d_j is also *design consistent*; i.e.,

$$\text{plim}_{\pi}(d_j - \bar{y}_{jP}) = 0, \\ n_j \rightarrow \infty$$

where π denote the probability space generated by the random selection process rather than the model in (1).

Isaki and Fuller (1982) give sufficient conditions for d_j to be design consistent, and it is under most sampling designs in common practice. Notable exceptions involve systematic sampling from a predetermined list (see Kott 1986). A popular alternative to design consistency is Brewer's (1979) *asymptotic design unbiasedness* (ADU) property. The estimator d_j is always ADU.

The trouble with d_j is that it may not be very efficient for small n_j . One solution is to “draw strength” from the other domains by treating the fixed parameter θ_j as if it was a realization of a random variable satisfying this *linking* model:

$$\theta_j = \mu + \tau_j, \quad (3)$$

where $E(\tau_j) = 0$, and $E(\tau_j \tau_g) = \sigma^2$ when $j = g$ and 0 otherwise. This is sometimes called “random effects modeling,” because the heretofore fixed effect of being a unit in domain j , θ_j , is now being treated as a random variable.

Combining equations (1) and (3) results in the reduced form components-of-variance model:

$$y_{ji} = \mu + \tau_j + \epsilon_{ji}. \quad (4)$$

Many analysts start with equation (4). We have separated the basic and linking models to underscore the greater level of confidence one often has in the validity of the basic model (especially when it is assumed as part of the linking model that all $\delta_g^2 = \delta^2$, as it soon will be).

Any estimator of the form:

$$f_j(\alpha, c) = (1 - \alpha)d_j + \alpha \hat{\mu},$$

where

$$c = (c_1, \dots, c_{j-1}, 0, c_{j+1}, \dots, c_m),$$

$$\hat{\mu} = \sum_{g=1}^m c_g \bar{y}_{gS},$$

$$\bar{y}_{gS} = \sum_{i=1}^{n_g} y_{gi} / n_g,$$

and

$$\sum_{g=1}^m c_g = 1$$

is unbiased under the model in (4). (Note: although the variables c and $\hat{\mu}$ depend on domain j , additional denotation has been suppressed for simplicity.)

If all the δ_g^2 are assumed equal to δ^2 , then using a Lagrangian multiplier technique it is not difficult to show that the choices for α and the c_g that minimize the model variance of $f_j(\alpha, c) - \bar{y}_{jP}$ are

$$\alpha^* = \frac{\sum_{i=1}^{n_j} w_{ji}^2 - 1/N_j}{\sum_i w_{ji}^2 + \sum_g c_g^{*2} / n_g + (1 + \sum_g c_g^{*2}) (\sigma^2 / \delta^2)}, \quad (5)$$

and

$$c_g^* = \frac{[(\sigma^2/\delta^2) + n_g^{-1}]^{-1}}{\sum_h [(\sigma^2/\delta^2) + n_h^{-1}]^{-1}}, \quad \text{for } g \neq j. \quad (6)$$

In practice, σ^2 and δ^2 are rarely known. Ghosh and Meeden (1986) have proposed estimating the ratio σ^2/δ^2 from the sample in a *model* consistent manner (as $m \rightarrow \infty$) by

$$L = \max \left\{ 0, \left[\frac{\sum_g n_g (\bar{y}_{gS} - \bar{y}_S)^2 / (m - 1)}{\sum_g \sum_i (y_{gi} - \bar{y}_{gS})^2 / (n - m)} - 1 \right] (m - 1) / (n - \sum_g n_g^2 / n) \right\}, \quad (7)$$

where

$$\bar{y}_S = \sum n_g \bar{y}_{gS} / n$$

and

$$n = \sum n_g.$$

Let $\alpha'(L)$ and $c'(L)$ be the right hand sides of equations (5) and (6) respectively with L replacing σ^2/δ^2 . Now call

$$e_j = f_j[\alpha'(L), c'(L)]$$

the *random effects estimator*, where $\hat{\mu}$ in $e_j = f_j(.,.)$ is set equal to $\mu'(L) = \sum c'_g(L) \bar{y}_{gS}$. As m grows large, e_j become indistinguishable from $f_j(\alpha^*, c^*)$.

If the model in (4) is correct and all the $\delta_j^2 = \delta^2 > 0$, then for sufficiently large m , L must be positive. Even if the model fails, as long as L is bounded from below by a positive number, $|\mu'(L)|$ is bounded, and $n_j \sum_{i=1}^n w_{ji}^2$ is bounded as n_j (but not m) grows arbitrarily large, then e_j is design consistent whenever d_j is. This is because

$$\text{plim}_{\pi} [\alpha'(L)] = 0,$$

so that e_j converges to the design consistent d_j .

3. MODEL AND DESIGN MEAN SQUARED ERROR

Under some sampling designs there exists an estimator of the design variance of d_j that is also a model unbiased estimator of the variance of d_j as an estimator for \bar{y}_{jP} under the basic model (henceforth I will omit the clarifying phrase “as an estimator for \bar{y}_{jP} ” to simplify the

exposition). Often, however, one must settle for a design consistent estimator of the design mean squared error of d_j (assuming, as we will, one exists). This is particularly true when $\sum_{k=1}^{n_j} p_{jk}^{-1} \neq N_j$. Kott (1987) shows how (when necessary) this estimator of the design mean squared error of d_j can be adjusted to be simultaneously a design consistent estimator of the design mean squared error of d_j and a model unbiased estimator of the variance of d_j under the *basic* model. Call this adjusted "variance estimator" $v(d_j)$.

We are now ready to address the model and design mean squared errors of the random effects estimator, e_j . Although we needed to assume that the δ_j^2 were all equal to determine e_j , we need not make that assumption in assessing the accuracy of e_j . In fact, we need not even assume that the linking model in equation (3) holds! Instead, we assume only that m is large enough so that L may be viewed as (virtually) independent of the units in domain j . Alternatively, L can be redefined by excluding units from domain j in the summations on the right hand side of (7).

Either way, $E_\epsilon[(d_j - \bar{y}_{jP})(\bar{y}_{jP} - \mu'(L))] = 0$. As a result,

$$E_\epsilon[\{d_j - \mu'(L)\}^2] = \text{var}_\epsilon(d_j - \bar{y}_{jP}) + E_\epsilon[\{\bar{y}_{jP} - \mu'(L)\}^2].$$

It is now a simple matter to show that under the basic model in (1),

$$v(e_j) = [1 - 2\alpha'(L)] v(d_j) + [\alpha'(L)]^2 [d_j - \mu'(L)]^2$$

is an unbiased estimator of the model mean squared error of e_j given L and $\mu'(L)$. Since $\alpha'(L)$ is asymptotically zero as n_j approaches infinity, $v(e_j)$ is also a design consistent estimator of the design mean squared error of e_j whenever $v(d_j)$ is a design consistent estimator of the design mean squared error of d_j .

It is not necessary for L to converge to σ^2/δ^2 or $\mu'(L)$ to converge to μ for $v(e_j)$ to have the properties described above. In fact, it is not necessary for the limits of L and $\mu'(L)$ to have any interpretations at all, since these properties have been defined independently of the model in equation (3).

Statisticians often have much more confidence in the basic model in equation (1) than the linking model in equation (3), especially when the latter is coupled with the assumption of constant unit variances (δ_g) across domains. It is therefore reassuring that the accuracy of the e_j can be estimated without invoking (3) or requiring that the δ_g be equal.

Unfortunately, $v(e_j)$ is unstable and can even be negative when $\alpha'(L)$ exceeds 0.5. Nevertheless, a simple comparison of the relative sizes of $v(d_j)$ and $v(e_j)$ over the m domains ($j = 1, \dots, m$) provides a robust method for choosing between the two estimators, d_j and e_j .

4. AN EMPIRICAL EXAMPLE

The Human Nutrition Information Service (HNIS) conducted a stratified, multistage survey of one day food intake by women aged 19-50 in 1985 as part of its Continuing Survey of Food Intakes by Individuals (CSFII). Responses were converted into measured intakes from among 60 food groups and 27 nutrients. See Human Nutrition Information Service (1985) for more details on the survey and its sample design.

We will restrict our attention here to the estimation of mean intake of milk and milk products (one of the 60 food groups) by women 19-34 and 35-50 within 12 mutually exclusive domains. These domains are defined by two cross classifications: region (northeast, midwest, south, and west) and level of urbanization (central city, suburban, non-metropolitan). HNIS published mean food group intakes separately for these two age groups on the national level only. Mean nutrition intakes were published for each age group by region and level of urbanization but were not cross-classified.

The CSFII sample design employed an independent stratified multistage sample with each of these domains. First primary sampling units (cities or town) were chosen using probability proportional to size sampling *with* replacement, then a random subsample of area segments was selected from which a smaller random subsample of households were chosen. I added another level of subsampling. When more than one woman per household from an age group was in the CSFII sample, I randomly chose one.

For each group, d_j in equation (2) defines the conventional design-based estimated of the domain mean. The SESUDAAN program (Shah 1980) provided design consistent estimators of all the d_j and their design root mean squared errors ($\sqrt{\text{MSE}(d_j)}$). These estimators, when squared, are not necessarily model unbiased estimators of the model variance of d_j under equation (1) however.

To see this, we confine our attention not only to an age group but to a domain as well and suppress the subscript j . Let $h = 1, \dots, H$ denote strata, $k = 1, \dots, K_h$ denote primary sampling units (PSU's) in h , and $i = 1, \dots, n_{hk}$ denote sampled women in hk . The estimate for the mean intake estimate is

$$d = \sum_{h=1}^H \sum_{k=1}^{K_h} \sum_{i=1}^{n_{hk}} w_{hki} y_{hki}.$$

We need more notation before we proceed. Let

$$x_{hk} = \sum_{i=1}^{n_{hk}} w_{hki},$$

$$z_{hk} = \sum_{i=1}^{n_{hk}} w_{hki}^2,$$

$$f_{hk} = \sum_{i=1}^{n_{hk}} w_{hki} (y_{hki} - d),$$

and

$$f_h = \sum_{k=1}^{K_h} f_{hk} / K_h.$$

If we assume the population size of the domain is large enough to be ignored (this also virtually assures that no individual had been sampled twice), the model variance of d is

$$\begin{aligned}\text{var}_\epsilon(d) &= \delta^2 \sum_h \sum_k \sum_i w_{hki}^2 \\ &= \delta^2 \sum_h \sum_k z_{hk}.\end{aligned}$$

The SESUDAAN (linearization) estimator for the design mean squared error of d is

$$v^*(d) = \sum_{h=1}^H (K_h/[K_h - 1]) \sum_{k=1}^{K_h} (f_{hk} - f_h)^2.$$

After much manipulation the model expectation of this can be shown to be

$$\begin{aligned}E_\epsilon[v^*(d)] &= \delta^2 \left[\sum_h \sum_k z_{hk} \right. \\ &\quad - 2 \sum_h (K_h/[K_h - 1]) \left(\sum_k^{K_h} z_{hk} x_{hk} - \sum_k^{K_h} z_{hk} \sum_k^{K_h} x_{hk}/K_h \right) \\ &\quad \left. + \left(\sum_h \sum_k z_{hk} \right) \sum_h (K_h/[K_h - 1]) \left(\sum_k^{K_h} x_{hk}^2 - \left\{ \sum_k^{K_h} x_{hk} \right\}^2 / K_h \right) \right].\end{aligned}$$

Following Kott (1987),

$$v(d) = v^*(d) \text{var}_\epsilon(d) / E_\epsilon[v^*(d)]$$

is both a design consistent estimator for the mean squared error of d (under certain conditions) and a model unbiased estimator of the model variance of d .

Calculations for n_j , d_j , $\alpha'(L)$, e_j , $v(d_j)$ and $v(e_j)$ for the 12 domains in each of the two groups are displayed in Table 1 (the domain subscript j has been returned to d_j and e_j). Using equation (5), L was calculated to be 0.055 for women 19-34 and 0.037 for women 35-50. This suggests that women in the same domain had little in common over and above their membership in the same age group. Nevertheless, $\alpha'(L)$ exceeded 0.5 only for five (out of 24) cells all with samples of under 25 women.

The estimate $v(e_j)$ was negative twice and less than $v(d_j)$ 18 out of 24 times, nine times for each age group. These latter group of numbers suggest to me that the e_j are indeed better estimates than the d_j . Formally, if we treat each of the 24 differences, $v(e_j) - v(d_j)$, as if they were independent across domains (they aren't quite), the hypothesis that the true model (or design) mean squared errors of e_j and d_j are equal and the random variable $v(e_j) - v(d_j)$ as likely positive as negative is soundly rejected.

The reduction in mean squared error from using e_j in place of d_j is estimated (by $\sum \{v(e_j) - v(d_j)\} / \sum v(d_j)$) to be 40.6%. This translates into a standard error reduction of 22.9%. Note that because we are summing 24 near independent random variates, we have much more confidence in this estimate than any particular $v(e_j)$ (or $v(d_j)$ for that matter).

Table 1
Estimated Values for the Domains by Age Group

Domain	Women 19-34					
	Sample Size	d_j	e_j	$v(d_j)$	$v(e_j)$	$\alpha'(L)$
N - C	68	220.6	222.1	683.0	367.5	.233
N - S	95	195.7	203.1	568.8	367.8	.225
N - R	12	219.1	223.8	5266.7	-1349.5	.630
M - C	55	270.7	258.6	2021.5	1152.5	.251
M - S	107	277.2	267.8	625.8	509.6	.164
M - R	73	301.1	285.9	4027.1	2754.3	.187
So - C	66	212.4	215.7	3011.6	1700.1	.220
So - S	112	156.8	167.9	472.8	457.3	.146
So - R	81	117.0	139.3	592.0	868.9	.184
W - C	39	403.0	333.2	2064.2	5438.4	.364
W - S	74	205.0	209.6	1704.0	1018.3	.207
W - R	13	120.0	190.7	3533.5	3924.3	.652
Women 35-50						
N - C	44	205.3	197.4	1716.1	318.4	.425
N - S	67	135.0	153.1	1068.8	698.0	.326
N - R	21	206.1	195.4	579.2	56.6	.550
M - C	28	89.0	139.5	470.3	2559.9	.482
M - S	87	200.3	196.1	2128.5	1049.2	.258
M - R	38	304.9	250.7	6065.3	3973.9	.415
So - C	47	136.1	159.6	266.7	592.6	.421
So - S	93	161.0	167.7	1492.5	809.1	.244
So - R	77	128.8	146.3	1023.4	790.9	.263
W - C	23	205.5	193.9	7497.1	-1067.6	.580
W - S	88	245.1	229.1	2484.7	1432.2	.263
W - R	11	132.1	173.3	743.3	1344.1	.734

Domain Codes
N - Northeast; M - Midwest; So - South; W - West; C - Central City; S - Suburban; R - Non-metropolitan.

5. DISCUSSION

Let $n_j^* = 1/\sum_{i=1}^{n_j} w_{ji}^2$ define the *effective sample size* within domain j . Observe that $n_j^* \leq n_j$ where equality holds if and only if all the sampling weights within j are all equal to $1/n_j$. For a known σ^2/δ^2 , the only difference between the optimal estimator developed here, $f_j(a^*, c^*)$, and the best linear unbiased predictor in Scott and Smith (1969) is that $1/n_j^*$ has replaced $1/n_j$ in the formula for α^* (equation (5)). The effect of this when the w_{ji} within j are not all equal is to increase α^* ; that is, to increase the dependence on sample information from outside domain j . This happens because forcing the estimator to be design consistent results in the domain j sample not being used as efficiently as possible. We could penalize the sample from outside the domain in a conformal manner by using sample weights in determining $\mu'(L)$, but that would only decrease the model efficiency of the estimator without improving any design-based characteristic.

Equation (7) assures that L can be no less than zero. This means that $\alpha'(L)$ can be no greater than $\sum_{g \neq j} n_g / (\sum_{g \neq j} n_g + n_j^*)$. If $\alpha'(L)$ were equal to its upper bound and $n_j^* = n_j$, then e_j would collapse into the simple mean of the y_{gi} across the entire sample. This makes sense because when the full model in equation (4) is correct and $\sigma^2 = 0$, the most efficient estimator of $\mu + \tau_j = \mu$ is the full sample mean.

If $n_j^* < n_j$ and $L = 0$, however, then e_j will be calculated with more weight given to units outside of domain j than to units inside the domain, which makes little sense. One *ad hoc* way to get around this phenomenon is to set an upper bound of $1 - (n_j / \sum n_g)$ (or smaller) on $\alpha'(L)$. Another approach would be to abandon small domain estimation entirely when $\alpha'(L)$ as calculated in the text exceeds $1 - (n_j / \sum n_g)$. Note that L , the estimated value for σ^2 / δ^2 , would have to be very small for this to happen. In the empirical study discussed in the previous section, L was in the 0.03 to 0.06 range, yet $\alpha'(L)$ was always well below $1 - (n_j / \sum n_g)$.

There are two ways the full model in equation (4) may fail. The fixed effects model within each domain (equation (1)) can fail or the linking model in (3) can fail. In the real world, both models are likely to be wrong. Equation (1) for its part ignores stratification and clustering effects as well as any subtle effect of membership in a household with more than one woman in the same age group. None of these effects are likely to be great. Moreover, by incorporating sampling weights into the estimate d_j and forcing the mean squared error estimators to be design consistent, we have done as much as we can do to protect ourselves against the potential for model failure in equation (1).

On the other hand, we should have little faith in the viability of the linking model. It is hardly more than a statistical convenience that, among other things, fails to allow for any correlation in the intakes of women from the same region but from different levels of urbanization or *vice versa*.

As noted, simply counting the number of times $v(e_j) - v(d_j)$ is negative provides a means for choosing between the estimators d_j and e_j that is independent of the linking model. The estimator $v(e_j)$ is unstable, however, and should not be used by itself as an estimate of mean squared error in practice.

Not only are the estimates of the mean squared error of e_j unstable, the $v(d_j)$ are only slightly better. At best $v(d_j)$ has "degrees of freedom" equal to the number of PSU's minus the number strata in j . For the CSFII sample, these range from 2 to 7.

Since it is becoming increasingly necessary for statisticians to provide estimated standard errors along with the estimated means they publish, it is imperative that more stable estimators than $v(d_j)$ and $v(e_j)$ be found. One idea might be to fit the $v(d_j)$ and the $v(e_j)$, either together or separately, with a variance estimating function. This approach is *ad hoc*, however, and may do little more than return values close to fully model-dependent estimates of the mean squared errors of the d_j and e_j (see Prasad and Rao 1986, for a good discussion of these) by "averaging out" the effects of model failure.

One intriguing idea is to combine the stable, but biased, model-dependent mean squared error estimates with the design consistent estimates developed here, much like e_j does for means. How this should be done is a topic that deserves future attention.

ACKNOWLEDGEMENTS

The author would like to thank the Human Nutrition Information Service for permission to use their data and Joe Goldman for his assistance in constructing the data sets used in the empirical analysis. John Herbert and a pair of anonymous referees must also be acknowledged for their helpful comments on earlier drafts of the manuscript.

REFERENCES

- BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.
- BREWER, K. R. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5, 93-105.
- BREWER, K. R. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.
- FAY, R. E., and HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FULLER, W. A., and HARTER, R. M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*, (Eds. R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh). New York: John Wiley and Sons.
- GHOSH, M., and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.
- GONZALEZ, M.E., and HORA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.
- HAJEK, J. (1971). Comment. In *Foundations of Statistical Inference*, (Eds. V. P. Godambe and D. A. Sprott). Toronto: Holt, Rinehart, and Winston.
- HUMAN NUTRITION INFORMATION SERVICE (1985). *CSFII – Nationwide Food Consumption Survey Continuous Survey of Food Intake by Individuals: Women 19-50 Years and Their Children 1-5 Years, 1 Day*. NFCS, CSFII Report No. 85-1. Washington: United States Department of Agriculture.
- ISAKI, C. T., and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTT, P. S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, 73, 485-491.
- KOTT, P. S. (1987). Estimating the conditional variance of a design consistent regression estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 486-491.
- PRASAD, N. G. N., and RAO, J. N. K. (1986). On the estimation of mean square error of small area predictions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 108-116.
- SÄRNDAL, C. E. (1984). Design consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCOTT, A., and SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.
- SHAH, B. V. (1981). *SESUDAAN: Standard Error Program for Computing of Standardized Rates from Sample Survey Data*. Research Triangle Park: Research Triangle Institute.
- STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics*, (Eds. R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh). New York: John Wiley and Sons.

Estimation of Livestock Inventories Using Several Area- and Multiple-Frame Estimators

GEORGE E. BATTESE, NANCY A. HASABELNABY, and WAYNE A. FULLER¹

ABSTRACT

Estimation of total numbers of hogs and pigs, sows and gilts, and cattle and calves in a state is studied using data obtained in the June Enumerative Survey conducted by the National Agricultural Statistics Service of the U.S. Department of Agriculture. It is possible to construct six different estimators using the June Enumerative Survey data. Three estimators involve data from area samples and three estimators combine data from list-frame and area-frame surveys. A rotation sampling scheme is used for the area frame portion of the June Enumerative Survey. Using data from the five years, 1982 through 1986, covariances among the estimators for different years are estimated. A composite estimator is proposed for the livestock numbers. The composite estimator is obtained by a generalized least-squares regression of the vector of different yearly estimators on an appropriate set of dummy variables. The composite estimator is designed to yield estimates for livestock inventories that are "at the same level" as the official estimates made by the U.S. Department of Agriculture.

KEY WORDS: June Enumerative Surveys; Rotation sample; Composite estimator; Generalized least squares.

1. INTRODUCTION

The National Agricultural Statistics Service (NASS), formerly the Statistical Reporting Service, of the U.S. Department of Agriculture (USDA) conducts probability surveys in June each year (the June Enumerative Surveys) to obtain data on farming operations. The survey data are a critical input in the construction of the official estimates of livestock numbers, crop acreages, grain stocks, *etc.* for the different states and for the United States as a whole. The sampling units in the farm surveys are selected from area frames and from list frames.

The area frame for a given state is the geographic area of the state stratified according to land use. The strata are defined by the percentage of the area that is cultivated, and whether the area is mainly urban, woodland, lakes, or other nonagricultural land. The sampling units for the area samples are called "segments", which vary in size in different states and strata, but are approximately one square mile in rural areas.

For the estimation of livestock inventories, samples of farm operators are also drawn from lists of farmers who raise the particular livestock. These list frames are stratified by measures of size. The area-frame and list-frame survey data are combined to obtain multiple-frame estimators for livestock numbers at the state level.

Different estimators can be constructed from the area-frame and list-frame samples. Statisticians within a state office of NASS calculate several estimators and make a recommendation for the official estimate of the number of livestock in the state. These materials are sent to the Agricultural Statistics Board within NASS in Washington D.C. The Board considers the different sample estimators, the recommendations of the state office, industry data, regional-level summaries and balance sheets when constructing the official estimates. Charting techniques

¹ George E. Battese, Department of Econometrics, University of New England, Armidale, N.S.W. 2351 Australia. Nancy A. Hasabelnaby and Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011, United States.

to maintain historical relationships among the data sources are also used by the Board. The Agricultural Statistics Board sets official estimates so that the official state estimates sum to the official national estimate.

A major drawback of the present procedure of establishing the official estimate is that there is no statistical measure of precision available for the official estimate. In 1983, a long-range planning group within NASS recommended that an objective procedure be developed for combining the different probability-based estimators into a composite estimator for the official estimate [see Allen, *et al.* 1983]. In 1984 it was recommended that a composite estimator should be made available for the consideration of the Agricultural Statistics Board [see Bynum, *et al.* 1985, p. 2].

The pooling of data from different, but related, samples and the combining of two or more estimators has been a subject of statistical research for many years. Some of this research is cited by Kuo (1986). Kuo also considers a composite estimator for livestock inventories based on USDA survey data.

In this study we investigate a procedure for constructing a composite estimator for livestock numbers. The values of several estimators for livestock inventories for a number of years and the variances and covariances among estimators for the different years are used in the construction. Assuming that the relationships among these estimators are defined by a simple linear model, we obtain the generalized least-squares estimator for the livestock inventories in the last year of sample data. Because the time-series of estimates is important, the set of composite estimators is constrained such that the average of the estimates for all years prior to the current year is equal to the average of the corresponding official estimates. This preserves the level of the time-series relative to previous official estimates. Alternative level constraints could be imposed.

2. AREA- AND MULTIPLE-FRAME ESTIMATORS

In the area-frame June Enumerative Survey, sample segments are identified on maps and all farm operators who have farming activities within these segments are identified and interviewed. The interviewers determine whether or not the farm operators in a given sample segment have their residences located within that segment. An area (or a collection of areas) of land within a sample segment that is under one type of management arrangement is called a "tract". A tract may be part of a farm or an entire farm.

The interviewer obtains information on the farming operation for each tract within a sample segment, including the size of the tract. In addition, information is obtained on the total farming operation of each sample farm operator. This information can be used to construct three different estimators of totals. The three estimators are called the closed-, open- and weighted-segment area-frame estimators. They differ mainly in the way in which farm values are associated with the segment.

The closed-segment area-frame estimator uses values associated with the operation of each tract within a sample segment. The open-segment area-frame estimator uses the values for the entire farm operation for those farms whose operators have their residences within the sample segment. The weighted-segment area-frame estimator uses values for the entire farm operation for farms with tracts in the sample segment. The values are prorated to the tract level by multiplying the farm total by the proportion of the total farm area that is within the sample segment. The weighted-segment value for a segment is the sum of the prorated values summed over all tracts within the sample segment. The closed-, open- and weighted-segment area-frame

estimators of totals are defined by multiplying the corresponding segment values by their segment weights (inverses of the probabilities of selection of the segments) and adding these values over all sample segments and strata within the state. The three estimators are compared and discussed by Houseman (1975) and Nealon (1984).

The closed-segment area-frame estimator is considered to have a smaller variance than the open-segment area-frame estimator for most variables that can be easily reported on a tract basis. Items such as farm expenditures and livestock deaths are not easily reported at the tract level. The closed-segment area-frame estimator is preferred for estimation of national crop acreages and is also calculated, along with other estimators, for livestock inventories in most states. When values of variables can easily be associated with tracts, the closed-segment area-frame estimator is generally preferred because it is believed that the data obtained are less subject to reporting error by farm operators than information for the whole farm.

The weighted-segment area-frame estimator generally has the smallest variance of the three area-frame estimators. The weighted-segment estimator can be used for estimation of the population total for any agricultural item. Nealon (1984, p. 19) cites several research studies which show that the weighted-segment area-frame estimator is biased because the total farm size is frequently underreported. It is generally believed that some areas in woodland, pastureland, idle land, and farmsteads are not reported as part of the farm. If so, the ratio of the tract area to the total farm area will be too large and the weighted-segment area-frame estimator will be positively biased.

Multiple-frame estimators for livestock inventories use sample data from two or more frames. In the case of livestock, there are usually two frames, the area frame and a list frame. The list frame is a list of operators that were known, at one time, to have the livestock of interest. The list frame is incomplete but generally contains many of the large operators. For estimation of hog inventories in the study state, multiple-frame estimators are obtained by summing the estimator for the total of the list frame constructed with the list sample and an estimator for the nonoverlap domain (those operators not found in the list frame) from the area sample. The list sample is considered to be independent of the area sample. Different multiple-frame estimators are obtained when the closed-, open- and weighted-segment area-frame estimators are used for the nonoverlap domain.

3. COMPOSITE ESTIMATOR

We propose a composite estimator for the livestock inventory constructed under the assumption that a linear model defines the relationship among the different estimators. Suppose that N estimators for a given livestock inventory are available in each of T years and that the Agricultural Statistics Board has made official estimates for the first $T-1$ years. It is assumed that a composite estimator for the livestock inventory in the T -th year is desired.

Let Y_{ti} represent the i -th estimator for the t -th year, where $t = 1, 2, \dots, T$ and $i = 1, 2, \dots, N$. We assume the linear model,

$$Y_{ti} = \alpha_t + \beta_i + e_{ti}, \quad (3.1)$$

where α_t is the livestock inventory for the t -th year;

β_i is the effect associated with the i -th estimator; and

e_{ti} is a random error which has mean zero.

The estimator effects, $\beta_1, \beta_2, \dots, \beta_N$, are included to account for the fact that nonsampling errors may cause different estimators to have different expectations. Model (3.1) specifies the estimator effects to be additive and constant over years. The assumption of constant effects is a simple specification that is consonant with the data.

The model (3.1) is a classical two-way, analysis-of-variance model, whose parameters are not estimable without additional model assumptions. To identify the parameters of the model, we restrict the average of the true livestock inventories in the first ($T-1$) years to be equal to the average of the corresponding official estimates of the Agricultural Statistics Board. This restriction is

$$\sum_{t=1}^{T-1} \alpha_t = \sum_{t=1}^{T-1} a_t, \quad (3.2)$$

where a_t is the official estimate for the t -th year. This constraint forces the estimates of livestock inventories to be at the same level as the previous official estimates. This is judged a reasonable constraint because actual values for α_t cannot be obtained and the time series nature of the estimates is important.

Given the restriction (3.2), the linear model (3.1) can be expressed in terms of the parameters, $\alpha_2, \alpha_3, \dots, \alpha_T$ and $\beta_1, \beta_2, \dots, \beta_N$, as

$$Y_{1i}^* = - \sum_{j=2}^{T-1} \alpha_j + \beta_i + e_{1i} \quad (3.3)$$

$$Y_{ti} = \alpha_t + \beta_i + e_{ti}$$

where $t = 2, 3, \dots, T$; and $Y_{1i}^* \equiv Y_{1i} - \sum_{j=1}^{T-1} a_j, i = 1, 2, \dots, N$.

The model in matrix notation is

$$Y^* = X\gamma + e, \quad (3.4)$$

where $Y^* \equiv (Y_{11}^*, \dots, Y_{1N}^*, Y_{21}, \dots, Y_{2N}, \dots, Y_{T1}, \dots, Y_{TN})'$;

X is the $(NT \times K)$ matrix of dummy variables associated with the model (3.3), where $K = T - 1 + N$;

$\gamma \equiv (\alpha_2, \alpha_3, \dots, \alpha_T, \beta_1, \beta_2, \dots, \beta_N)'$; and

e is the NT -column vector of random errors, having covariance matrix, V .

The covariance matrix, V , is the covariance matrix of the sampling errors, e_{ti} , associated with the different estimation procedures. The estimators, Y_{ti} , $t = 1, 2, \dots, T$; $i = 1, 2, \dots, N$, are correlated within any given year because they are based on the same area segments and the same list sample. The estimators are also correlated among years because sample segments in the area sample are included in the surveys for several years, according to a rotation sampling scheme. The list sample is selected independently each year. The variances and

covariances of the estimators for any given year can be estimated by standard survey-sampling methods. Because the same list-sample estimator is used in defining the three multiple-frame estimators in a given year, the covariance between any two of the multiple-frame estimators in the same year will have a component due to the variance of the estimator obtained from the list sample. The covariances between estimators in different years, $Cov(Y_{ti}, Y_{t'j})$, where $t \neq t'$, can be estimated by standard methods, using the sample segments that are common to the two years. If it is assumed that the variances and covariances in V satisfy particular relationships, then these conditions can be imposed as part of the estimation procedure.

Given an estimator of the covariance matrix, denoted by \hat{V}^* , the estimated generalized least-squares estimator of the parameter vector, $\hat{\gamma}$, is

$$\hat{\gamma} = (X'\hat{V}^{*-1}X)^{-1}(X'\hat{V}^{*-1}Y^*). \tag{3.5}$$

The covariance matrix of $\hat{\gamma}$ is estimated by

$$Cov(\hat{\gamma}) = (X'\hat{V}^{*-1}X)^{-1}. \tag{3.6}$$

The estimated generalized least-squares estimator, $\hat{\alpha}_T$, which is the $(T-1)$ -th element of $\hat{\gamma}$, is a possible composite estimator for the livestock inventory for the T -th year. Its variance is estimated by the corresponding element of the estimated covariance matrix (3.6). Furthermore, the estimated generalized least-squares estimators, $\hat{\alpha}_T + \hat{\beta}_i, i = 1, 2, \dots, N$, are adjusted area-frame and multiple-frame estimators for livestock inventories in the T -th year which are based on the model (3.4). The variances of these adjusted estimators are estimated by obtaining the appropriate linear functions of the estimated covariance matrix (3.6).

If the model (3.4) is true and the random errors have a normal distribution, then the weighted sum of squares,

$$x^2 = (Y^* - X\hat{\gamma})'\hat{V}^{*-1}(Y^* - X\hat{\gamma}), \tag{3.7}$$

has a chi-square distribution with parameter, $NT - K$. Thus the weighted residual sum of squares obtained by using the estimated covariance matrix yields an approximate test of the adequacy of the model (3.1).

4. EMPIRICAL RESULTS

4.1 Introduction

In the USDA June Enumerative Surveys between 1982 and 1986, a total of 298 area segments were sampled in the study state. These segments were included in the June Enumerative Surveys according to a rotation sampling scheme in which approximately twenty percent of the segments are replaced each year. The actual replacement rate varies, but we construct estimators as if the rate was exactly twenty percent.

The area frame for the state consists of eleven strata: nine strata are agricultural land, with varying percentages cultivated; one stratum is agri-urban land; and one stratum consists of residential or commercial areas.

The list frame for hog producers in the study state consists of eleven strata, which are defined by the total number of hogs raised by the farm operators at a particular time. The strata are defined by operations with: no livestock, livestock but no hogs, 1-99 hogs, 100-199 hogs, . . . , more than 6,000 hogs. The list frame for cattle that is sampled in June contains very large operators. It was a small list of less than 500 operators in each of the study years. The cattle list is divided into four strata. Three strata are defined by the total number of cattle and calves, where the strata are between 1,000 and 2,999, between 3,000 and 9,999, and more than 10,000. The fourth stratum is composed of farm operators with at least 200 dairy cattle.

The total number of farm operators in the area sample of the June Enumerative Surveys averaged about 2,350 during the years studied with a range of 120. The list sample for hogs averaged about 2,400 farm operators with a range of 100, whereas the list sample for cattle averaged about 70 farm operators with a range of 71. Using these data, the values of the closed-, open-, and weighted-segment area-frame estimators and the three corresponding multiple-frame estimators for the total number of hogs and pigs, sows and gilts, and cattle and calves were computed for each of the five years. The estimates were obtained by use of PC CARP, which is a computer program for performing survey-sampling estimation on personal computers [see Fuller, *et al.* 1986 and Schnell, *et al.* 1988]. The variance estimators are the usual estimators for an estimated total constructed from a stratified cluster sample. See, for example, Cochran (1977).

The data used for variance computations were treated as complete data although some data were imputed for nonresponse. The imputation, especially since the imputation methods draw heavily upon prior year data in the rotation scheme, may lead to an overestimate of the correlation between years.

Table 1
Estimates for livestock inventories in 1986

	Hogs and pigs	Sows and gilts	Cattle and calves
Area-Frame Estimators			
Closed-Segment	18.42 (1.97)	15.78 (2.17)	15.27 (1.53)
Open-Segment	21.11 (2.82)	18.24 (2.69)	18.74 (2.35)
Weighted-Segment	21.69 (1.67)	18.85 (1.62)	15.48 (1.15)
Multiple-Frame Estimators			
Closed-Segment	18.11 (1.11)	15.59 (1.28)	16.12 (1.38)
Open-Segment	18.06 (1.26)	15.29 (1.39)	19.97 (2.08)
Weighted-Segment	18.50 (1.00)	15.82 (1.00)	16.22 (1.00)

Estimates for the livestock inventories in 1986 and the estimated standard deviations of the estimators for 1986 are given in Table 1. Each standard deviation in Table 1 is the square root of the average of the five estimated variances for the five years. The units in the table are determined by coding the standard deviation of the weighted-segment multiple-frame estimators to be 1.00 for all livestock inventories. This makes comparison easy and also complies with confidentiality rules.

As expected from previous studies [*e.g.*, see Nealon (1984)], the open-segment area-frame estimator is the least precise estimator for livestock inventories. The most precise estimator is the multiple-frame estimator which uses the weighted-segment estimator for the nonoverlap domain. Coefficients of variation for the weighted-segment area-frame estimators are about 7% to 9%, whereas the weighted-segment multiple-frame estimators have coefficients of variation of about 5.5% to 6.5%. Because the list sample for hog inventories is larger than that for cattle and calves, the precision of the multiple-frame estimators relative to the area-frame estimators is much greater for hog inventories than for cattle and calves.

4.2 Estimation of Covariance Matrices

The estimation of the covariance matrix for the six estimators for the five years of data proceeded in several steps. The covariance matrix for the error vector, *e*, in (3.4) can be written in the form

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} & V_{14} & V_{15} \\ V_{21} & V_{22} & V_{23} & V_{24} & V_{25} \\ V_{31} & V_{32} & V_{33} & V_{34} & V_{35} \\ V_{41} & V_{42} & V_{43} & V_{44} & V_{45} \\ V_{51} & V_{52} & V_{53} & V_{54} & V_{55} \end{pmatrix}$$

(4.1)

where, for a particular inventory type, *V_{tj}* is the 6 x 6 matrix of covariances between the six estimators for year *t* and the six estimators for year *j*. With the rotation scheme used, the covariance of estimators in two years is a function of the number of rotation groups common to the two years. Let *k* = |*t* - *j*| for *k* = 0, 1, . . . , 4. Then the covariance matrix, *V_{tj}*, can be estimated from the area segments of the 5 - *k* rotation groups which are common to the two years *t* and *j*.

We estimate the elements of the covariance matrix (4.1) imposing some additional assumptions about its structure. Our primary interest is to compare the precision of the alternative estimators and this comparison is facilitated by the assumptions which follow.

We assume that the covariance matrices for years that are the same distance apart are the same and are symmetric. This is, we assume

$$V_{tj} = V_{t+r,j+r}$$

and

(4.2)

$$V_{tj} = V_{tj}'$$

where V_{ij} are the submatrices of (4.1); $r = 0, 1, \dots, \max(5-t, 5-j)$; and $t, j = 1, 2, \dots, 5$. For $t = j$ and $r = 0$, the assumptions of (4.2) imply

$$V_{11} = V_{22} = V_{33} = V_{44} = V_{55} \equiv V_0.$$

For $t \neq j$, the assumptions of (4.2) imply the following:

$$\begin{aligned} V_{12} &= V_{23} = V_{34} = V_{45} \equiv V_1, \\ V_{13} &= V_{24} = V_{35} \equiv V_2, \\ V_{14} &= V_{25} \equiv V_3, \end{aligned}$$

and

$$V_{15} \equiv V_4.$$

These assumptions are in reasonable agreement with the data. Good agreement was anticipated because the sample size is very stable over the five years and there were no large shifts in livestock inventories.

We estimate the distinct submatrices of (4.1) by averaging the corresponding estimated covariance matrices obtained from common segments. The averaging process was based on the correlation matrices. Let the covariance matrix of the estimated totals defined in (4.1) be expressed as

$$V = S C S,$$

where S is the 30×30 diagonal matrix of the estimated standard deviations of the six estimators for the five years and C is the 30×30 correlation matrix, partitioned in the same manner as V of (4.1).

The estimator of the correlation matrix C is constructed by averaging estimates of the submatrices of C . Using the segments common to two years, the covariance matrix of the two vectors of estimated totals constructed with those segments was estimated by the usual stratified cluster formulae. The estimated covariance matrices were converted to correlation matrices and these estimates were called the direct estimates. Let

$$\begin{aligned} \hat{C}_0 &= \left(\frac{5}{5}\right)\frac{1}{5}(\hat{C}_{11} + \hat{C}_{22} + \hat{C}_{33} + \hat{C}_{44} + \hat{C}_{55}) \\ \hat{C}_1 &= \left(\frac{4}{5}\right)\frac{1}{4}(\hat{C}_{12} + \hat{C}_{23} + \hat{C}_{34} + \hat{C}_{45}) \\ \hat{C}_2 &= \left(\frac{3}{5}\right)\frac{1}{3}(\hat{C}_{13} + \hat{C}_{24} + \hat{C}_{35}) \\ \hat{C}_3 &= \left(\frac{2}{5}\right)\frac{1}{2}(\hat{C}_{14} + \hat{C}_{25}) \\ \hat{C}_4 &= \left(\frac{1}{5}\right)\hat{C}_{15}, \end{aligned}$$

where the \hat{C}_{ij} are the directly estimated correlation matrices based on common segments. The factors in parentheses represent the fraction of segments that are common to the estimates. This fraction arises from the rotation-sampling scheme in which twenty percent of the segments in the area sample are dropped from the sample each year and twenty percent new segments are added. By the independence assumption, the correlation between the segments rotated out and those rotated in is zero.

Since the estimated correlation matrices, \hat{C}_{ij} , are not symmetric when $t \neq j$, the symmetric assumption, $V_{ij} = V'_{ij}$, in (4.2) is imposed on the estimated covariance matrix by defining

$$\hat{C}_r^* = \frac{1}{2}(\hat{C}_r + \hat{C}'_r), \quad r = 1, 2, 3, 4.$$

Let \hat{S}^* be the 6 x 6 diagonal matrix of the square roots of the average estimated variances of the six estimators, where the average is over the five years. Again, for confidentiality requirements, the estimated variances are standardized such that the estimated variance of the weighted-segment multiple-frame estimator is equal to 1.00. Then the estimated covariance matrix for the six estimators for the five years is

$$\hat{V}^* = \begin{pmatrix} \hat{S}^* & 0 & 0 & 0 & 0 \\ 0 & \hat{S}^* & 0 & 0 & 0 \\ 0 & 0 & \hat{S}^* & 0 & 0 \\ 0 & 0 & 0 & \hat{S}^* & 0 \\ 0 & 0 & 0 & 0 & \hat{S}^* \end{pmatrix} \begin{pmatrix} \hat{C}_0 & \hat{C}_1^* & \hat{C}_2^* & \hat{C}_3^* & \hat{C}_4^* \\ \hat{C}_1^* & \hat{C}_0 & \hat{C}_1^* & \hat{C}_2^* & \hat{C}_3^* \\ \hat{C}_2^* & \hat{C}_1^* & \hat{C}_0 & \hat{C}_1^* & \hat{C}_2^* \\ \hat{C}_3^* & \hat{C}_2^* & \hat{C}_1^* & \hat{C}_0 & \hat{C}_1^* \\ \hat{C}_4^* & \hat{C}_3^* & \hat{C}_2^* & \hat{C}_1^* & \hat{C}_0 \end{pmatrix} \begin{pmatrix} \hat{S}^* & 0 & 0 & 0 & 0 \\ 0 & \hat{S}^* & 0 & 0 & 0 \\ 0 & 0 & \hat{S}^* & 0 & 0 \\ 0 & 0 & 0 & \hat{S}^* & 0 \\ 0 & 0 & 0 & 0 & \hat{S}^* \end{pmatrix}$$

(4.3)

The estimated covariance matrices, $\hat{V}_0^* \equiv \hat{S}^* \hat{C}_0 \hat{S}^*$, for the livestock inventories are given in Table 2. The estimates of the four unique off-diagonal submatrices, $\hat{V}_r^* \equiv \hat{S}^* \hat{C}_r \hat{S}^*$, $r = 1, 2, 3, 4$, are available from the authors on request.

Table 2
Estimated covariance matrices for the six
estimators of livestock inventories within a year

	Area-Frame Estimators			Multiple-Frame Estimators		
	Closed	Open	Weighted	Closed	Open	Weighted
A. Hogs and pigs						
	3.886	4.077	2.366	0.654	0.688	0.405
	4.077	7.959	2.394	0.698	1.150	0.430
	2.366	2.394	2.784	0.373	0.409	0.481
	0.654	0.698	0.373	1.242	1.239	0.936
	0.688	1.150	0.409	1.239	1.590	0.937
	0.405	0.430	0.481	0.936	0.937	1.000
B. Sows and gilts						
	4.720	4.274	2.455	1.102	1.112	0.572
	4.274	7.260	2.322	1.119	1.427	0.548
	2.455	2.322	2.621	0.481	0.487	0.499
	1.102	1.119	0.481	1.638	1.658	1.033
	1.112	1.427	0.487	1.658	1.934	1.033
	0.572	0.548	0.499	1.033	1.033	1.000
C. Cattle and calves						
	2.355	1.951	1.141	1.853	1.655	0.907
	1.951	5.527	1.014	1.652	4.418	0.912
	1.141	1.014	1.321	0.913	0.891	0.925
	1.853	1.652	0.913	1.910	1.756	0.992
	1.655	4.418	0.891	1.756	4.310	1.017
	0.907	0.912	0.925	0.992	1.017	1.000

Consider a sample composed of a common set of rotation groups observed in each of the five years, rather than the existing sample in which twenty percent of the sample segments are dropped each year. For the sample with no rotation, the covariance matrix of the six estimators for the five years, expressed in terms of the submatrices of (4.1), is

$$\begin{pmatrix} V_{11} & \frac{5}{4} V_{12} & \frac{5}{3} V_{13} & \frac{5}{2} V_{14} & 5 V_{15} \\ \frac{5}{4} V_{21} & V_{22} & \frac{5}{4} V_{23} & \frac{5}{3} V_{24} & \frac{5}{2} V_{25} \\ \frac{5}{3} V_{31} & \frac{5}{4} V_{32} & V_{33} & \frac{5}{4} V_{34} & \frac{5}{3} V_{35} \\ \frac{5}{2} V_{41} & \frac{5}{3} V_{42} & \frac{5}{4} V_{43} & V_{44} & \frac{5}{4} V_{45} \\ 5 V_{51} & \frac{5}{2} V_{52} & \frac{5}{3} V_{53} & \frac{5}{4} V_{54} & V_{55} \end{pmatrix} \tag{4.4}$$

Direct sample estimates of the submatrices, V_{tj} , obtained from segments common to years t and j sometimes gave a covariance matrix (4.4) that was not positive definite. For example, this can happen if operators with very large holdings are among those operators in the one rotation common to all five years. When the assumptions of (4.2) are imposed in the estimation process, the estimates of the covariance matrix (4.4) were positive definite for all three livestock inventories.

Table 3
Composite estimates for the livestock inventories in 1986 and the effects for different estimators

	Hogs and pigs ¹	Sows and gilts ¹	Cattle and calves ¹
Composite Estimator	18.84 (1.01)	18.06 (1.02)	16.43 (1.03)
Effects of Area-Frame Estimators			
Closed-Segment	-1.13 (1.30)	-2.26 (1.36)	-0.21 (0.99)
Open-Segment	0.26 (1.86)	-1.09 (1.78)	1.03 (1.45)
Weighted-Segment	1.24 (1.14)	-0.94 (1.10)	-0.26 (0.80)
Effects of Multiple-Frame Estimators			
Closed-Segment	-0.33 (0.66)	-1.86 (0.78)	0.04 (0.92)
Open-Segment	-0.11 (0.75)	-1.82 (0.84)	1.40 (1.32)
Weighted-Segment	0.19 (0.59)	-1.74 (0.59)	-0.31 (0.69)

¹ Standard errors are given in parentheses.

4.3 Model Estimation

Given the estimated covariance matrix, \hat{V}^* , we estimate the parameters of model (3.4) by using the estimated generalized-least-squares estimator (3.5). The values of the composite estimator, $\hat{\alpha}_T$, for 1986 livestock inventories are given in the first line of Table 3. The six estimator effects, denoted by β_i in model (3.3), are also given in the table. The estimated standard deviation of the composite estimator is slightly larger than that of the weighted-segment multiple-frame estimator. The increase in variance comes from the fact that the level of the estimator is estimated using past sample estimates and past official estimates.

The residual sums of squares defined in (3.7) were 18.22, 15.38, and 24.59, for hogs and pigs, sows and gilts, and cattle and calves, respectively. The degrees of freedom is 20 because, for each livestock inventory, there are thirty observations in the Y^* -vector and ten parameters are estimated in γ . In no case does the residual sum of squares exceed 31.41, which is the 95-th percentile for the chi-square distribution with 20 degrees of freedom.

The composite estimator in Table 3 has nearly the same standard deviation as the weighted-segment multiple-frame estimator, the estimator with the smallest standard deviation (Table 1). Thus, one would expect the optimal linear combination of the six estimators for a single year to assign the majority of the weight to the weighted-segment multiple-frame estimator, and this is the case. The minimum variance weights for the data of a single year are calculated as

$$(\mathbf{1}'\hat{V}_0^{-1}\mathbf{1})^{-1}\mathbf{1}'\hat{V}_0^{-1},$$

where $\mathbf{1}' = (1, 1, 1, 1, 1, 1)$ and \hat{V}_0^* is the covariance matrix of the six estimators given in Table 2 [see the diagonal elements of (4.3)]. The optimal weights and the estimated standard deviation of the optimal combination of the six estimators are presented in Table 4. Note that the sum of the weights is one for each livestock inventory. The difference between these standard errors and those of the first line of Table 3 is due to the estimation of level in the construction of the estimates of Table 3.

Table 4
Optimal weights for six estimators in a single year.

Estimators	Inventory type		
	Hogs and pigs	Sows and gilts	Cattle and calves
Area-Frame			
Closed-Segment	0.0541	-0.0152	0.0525
Open-Segment	-0.0084	0.0152	0.0656
Weighted-Segment	0.1463	0.1909	0.0909
Multiple-Frame			
Closed-Segment	0.1640	-0.0218	-0.0353
Open-Segment	-0.0116	-0.0191	-0.0772
Weighted-Segment	0.6556	0.8500	0.9035
Estimated standard error of optimal combination	0.94	0.95	0.99

Table 5
Estimated correlation coefficients between the weighted-segment
area-frame estimators based on a common rotation group

h	Hogs & pigs	Sows & gilts	Cattle & calves
0	1.000	1.000	1.000
1	0.606	0.590	0.592
2	0.478	0.456	0.433
3	0.365	0.336	0.258
4	0.304	0.217	0.097

4.4 Estimation Using Rotation Group Means

In obtaining the estimates of Section 4.3, we did not use all of the available information. We used the estimators for each year, but did not decompose the estimators into the parts associated with each rotation group. In this section we construct an estimator using the individual rotation group means of the weighted-segment area-frame estimator. We retain the assumption that the variance of the estimator is the same across years. Under that assumption, the correlation coefficients are assumed to depend only on the number of years between the estimators involved. Let ρ_h represent the correlation coefficient between the weighted-segment area-frame estimators for a common rotation group which is observed h years apart, $h = 0, 1, \dots, 4$. For the three inventory types, the estimated correlation coefficients are given in Table 5. The estimated correlation coefficients between estimators for h years apart are the averages of the correlation coefficients estimated from the $5 - h$ rotation groups involved. There are a total of nine rotation groups for the five years.

Let Z_{tj} represent the weighted-segment area-frame estimator in rotation group j for year t , where $j = t, t + 1, \dots, t + 4$ and $t = 1, 2, \dots, 5$. Then, for a given year, t , we assume that Z_{tj} is an unbiased estimator of the unknown total inventory, α_t . It is known that a rotation group bias may exist and need to be estimated, but we ignore that effect in this illustration. The model is

$$Z_{tj} = \alpha_t + \epsilon_{tj}, t = 1, 2, \dots, 5;$$
$$j = t, t + 1, \dots, t + 4,$$

(4.5)

where the errors, ϵ_{tj} , have zero mean. The model (4.5), in matrix notation, is

$$Z = D\alpha + \underline{\epsilon},$$

where

$$Z' = (Z_{11}, Z_{12}, \dots, Z_{15}; Z_{22}, Z_{23}, \dots, Z_{26}; \dots;$$
$$Z_{55}, Z_{56}, \dots, Z_{59}),$$
$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_5),$$

$D = I_5 \otimes \mathbf{1}_5$, I_5 is the identity matrix of order 5 and $\mathbf{1}_5$ is the (5×1) vector with all elements equal to one.

Let the correlation matrix for the rotation-group estimators, Z , be

$$W = \begin{pmatrix} W_0 & W_1' & W_2' & W_3' & W_4' \\ W_1 & W_0 & W_1' & W_2' & W_3' \\ W_2 & W_1 & W_0 & W_1' & W_2' \\ W_3 & W_2 & W_1 & W_0 & W_1' \\ W_4 & W_3 & W_2 & W_1 & W_0 \end{pmatrix}$$

where $W_0 = I_5$,

$$W_1 = \begin{pmatrix} 0 & \rho_1 & 0 & 0 & 0 \\ 0 & 0 & \rho_1 & 0 & 0 \\ 0 & 0 & 0 & \rho_1 & 0 \\ 0 & 0 & 0 & 0 & \rho_1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0 & 0 & \rho_2 & 0 & 0 \\ 0 & 0 & 0 & \rho_2 & 0 \\ 0 & 0 & 0 & 0 & \rho_2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$W_3 = \begin{pmatrix} 0 & 0 & 0 & \rho_3 & 0 \\ 0 & 0 & 0 & 0 & \rho_3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad W_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & \rho_4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then, the generalized least-squares weighted-segment area-frame estimator is

$$\hat{\underline{\alpha}} = (D' \hat{W}^{-1} D)^{-1} D' \hat{W}^{-1} Z,$$

where \hat{W} is the estimator for the correlation matrix, W . The covariance matrix of $\hat{\underline{\alpha}}$ is estimated by

$$C\hat{ov}(\hat{\underline{\alpha}}) \equiv (D' \hat{\underline{\Sigma}}^{-1} D)^{-1},$$

where $\hat{\underline{\Sigma}} \equiv 5\hat{W}$ is the covariance matrix of Z , whose units are such that the estimated variance of the weighted-segment area-frame estimator is one. The estimated covariance matrices, $C\hat{ov}(\hat{\underline{\alpha}})$, for the three livestock types are given in Table 6. We see that the estimators obtained using the individual rotation group estimates are about 10% more efficient than the weighted-segment area-frame estimators for 1986.

The optimal weights for the vector of individual rotation estimates are

$$(D' \hat{W}^{-1} D)^{-1} D' \hat{W}^{-1}.$$

The weights are available from the authors.

The generalized least squares procedure can be applied to other combinations of rotation group and year estimators, but the results suggest that additional gains would be modest.

Table 6
Estimated covariance matrices for weighted-segment area-frame
estimators using information in the rotation scheme

	1982	1983	1984	1985	1986
Hogs and pigs					
1982	0.899	0.436	0.283	0.180	0.124
1983	0.436	0.857	0.412	0.273	0.180
1984	0.283	0.412	0.844	0.412	0.283
1985	0.180	0.273	0.412	0.857	0.436
1986	0.124	0.180	0.283	0.436	0.899
Sows and gilts					
1982	0.908	0.429	0.272	0.167	0.099
1983	0.429	0.866	0.405	0.262	0.167
1984	0.272	0.405	0.853	0.405	0.272
1985	0.167	0.262	0.405	0.866	0.429
1986	0.099	0.167	0.272	0.429	0.908
Cattle and calves					
1982	0.914	0.438	0.264	0.135	0.061
1983	0.438	0.870	0.412	0.253	0.135
1984	0.264	0.412	0.856	0.412	0.264
1985	0.135	0.253	0.412	0.870	0.438
1986	0.061	0.135	0.264	0.438	0.914

5. CONCLUSION

The composite estimator suggested in this paper provides a method for combining the values of several estimators for livestock inventories. The composite estimator uses the values of the different area-frame and multiple-frame estimators in several preceding years, as well as the values in the year for which the official estimate is sought. The optimal linear combination of the six estimators within a particular year has a variance that is two to twelve percent less than that of the weighted-segment multiple-frame estimator. Including the estimators from the other four years produces an additional reduction of one to two percent in the variance of the composite estimator for the current year. The data required to calculate the weighted-segment multiple-frame estimator are those required for the other five area- and multiple-frame estimators. The greatest effort required in constructing the composite estimator is the estimation of the covariance matrix for the estimators over the years in which sample data are available. Because the variances are relatively constant over years, the weight vector can be calculated in advance and applied to the estimates of the current year. Then the marginal effort required for the composite estimator during the estimation year is very small.

ACKNOWLEDGMENTS

This research was partially supported by Research Agreement No. 53-319T-6-00073 with the National Agricultural Statistics Service of the U.S. Department of Agriculture. The authors thank Ron Fecso and Vic Tolomeo for their assistance. Ron Fecso's comments on an earlier draft are gratefully acknowledged. The work was conducted when the first author was on study leave at Iowa State University.

REFERENCES

- ALLEN, R., CLAMPET, G., DUNKERLEY, C., TORTORA, R., and VOGEL, F. (1983). Framework for the Future. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- BYNUM, H., DOWDY, W., HANUSCHAK, G., HUDSON, C., MURPHY, R., STEINBERG, J., and VOGEL, F.A. (1985). Crop Reporting Board Standards. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- COCHRAN, W.G (1977). *Sampling Techniques*. New York: Wiley.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.
- HOUSEMAN, E.E. (1975). Area Frame Sampling in Agriculture. SRS Report No. 20. Statistical Reporting Service, U.S. Department of Agriculture, Washinton, D.C.
- KUO, L. (1986). Composite Estimation of Totals for Livestock Surveys. SF & SRB Staff Report No. 92. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. SF & SRB Staff Report No. 80. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology* 14, 59-69.

Modelling and Estimation for Repeated Surveys

D.A. BINDER and J.P. DICK¹

ABSTRACT

Estimation of the means of a characteristic for a population at different points in time, based on a series of repeated surveys, is briefly reviewed. By imposing a stochastic parametric model on these means, it is possible to estimate the parameters of the model and to obtain alternative estimators of the means themselves. We describe the case where the population means follow an autoregressive-moving average (ARMA) process and the survey errors can also be formulated as an ARMA process. An example using data from the Canadian Travel Survey is presented.

KEY WORDS: Kalman filter; Overlapping surveys; State-space models; Time series modelling; Small area estimates.

1. INTRODUCTION

When surveys with similar data items are conducted on repeated occasions, certain estimation and data analysis methods are available which are not possible with single occasion surveys. For example, efficient estimation methods for the current occasion can depend on data from previous occasions. This occurs when there are overlapping sampling units between occasions and, hence, the survey errors can be correlated over time. As well, the series of estimates from a repeated survey are often modelled by the data users. A common example of this is to assume an autoregressive-moving average (ARMA) model. However, most existing procedures for estimating the unknown parameters of this model assume that the input data are not subject to survey error.

In this paper we develop procedures for estimating these model parameters when the data contain survey errors. The covariance structure of the survey errors we consider include some cases where the survey errors are correlated over time.

When such a model for the behaviour of the population characteristics is assumed, the minimum mean squared error (MMSE) linear estimator can be derived. This estimator incorporates the model structure which the classical minimum variance linear unbiased estimator (MVLUE) ignores. The MVLUE is discussed in Section 2.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), R.G. Jones (1980) and others considered the implications of such stochastic models for the population means over time. These results and a more general formulation using state-space models and Kalman filters are discussed in Section 3, for the case where the stochastic model for the population characteristics is completely specified. These methods can be developed in a setting which is equivalent to a Bayes formulation, where the prior distribution is completely specified.

When the assumed model is an ARMA process in the presence of survey errors, the state-space formulation can be used to derive the maximum likelihood estimates of the unknown

¹ D.A.Binder and J.P. Dick, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

parameters. We note that this approach can be viewed as empirical Bayes. We assume that the survey errors can be described through an ARMA process up to a multiplicative factor. This is discussed in Section 4.

An example of this model is described in Section 5 using data from the Canadian Travel Survey. This example shows the implications on the estimates of the model parameters when the survey errors are taken into account. We also derive a smoothed estimate of the underlying process under the model assumptions. In this example, the survey errors are independent, so that the full machinery of the general formulation in this paper is not required. However, the example demonstrates that the impact of ignoring the survey errors even in this case can be appreciable.

Section 6 contains some concluding remarks.

2. MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION IN OVERLAPPING REPEATED SURVEYS

In this section we briefly review the literature for the case where the population values of a characteristic such as a mean or total are taken as fixed unknown constants. In Section 3, we study the case where a stochastic model is assumed for the population characteristic.

In overlapping surveys, where the same individual provides responses on repeated occasions, the sampling errors between occasions are usually correlated. Correlations can also occur in a multi-stage survey where some of the first stage sampling units overlap, even though the ultimate respondents differ.

Estimators which ignore these correlations and use only the data collected in the single reference period are in general inefficient relative to the minimum variance linear unbiased estimator (MVLUE). The relative efficiency depends on the size of the correlation of the sampling errors between occasions. When the correlations are zero, as in our example in Section 5, the MVLUE is simply the estimator based on data from a single reference period.

Jessen (1942) was the first to incorporate the overlapping information from the same individual on two successive occasions. Patterson (1950) provided a general theory for repeated surveys with overlapping units. He considered in detail the special case of simple random sampling from an infinite population, where the correlation for individuals is exponentially declining in time lag. On each occasions, a sample of individuals is removed from the sample of the previous occasion and a sample of individuals is added. All data are collected with reference to the current occasion only. Patterson derived the MVLUE for this setup.

Extensions have been made to the basic assumptions of Patterson (1950). Eckler (1955) called Patterson's design one-level rotation sampling. Eckler derived the MVLUE when individuals report for two successive time periods, which he termed two-level rotation sampling. He also derived the MVLUE for surveys with higher order rotation sampling designs.

Rao and Graham (1964) relaxed the infinite population assumption by incorporating the finite population correction factor into the variances of the survey error. Singh (1968) was the first to consider multi-stage designs. He examined two-stage sampling with the assumption that the correlation between responses on different occasions can be considered in two parts: (i) the correlation between second stage units (SSU's) within primary sampling units (PSU's) and (ii) the correlation between PSU means on successive occasions. If both of these correlation patterns are assumed to be that of a first order autoregressive process, then the form of the MVLUE follows the general form given by Patterson (1950).

Tikkiwal (1979) and others considered the implications of relaxing the assumption of a first order autoregressive correlation pattern. Tikkiwal concluded that if a completely general correlation structure is assumed, the simple form of the MVLUE is lost and approximations must be used in practice. Rao and Graham (1964) and Gurney and Daly (1965) proposed the use of composite estimators which are approximations to the optimal estimators. These estimators are easily implemented and have high relative efficiency. For a discussion on the use of these estimators, see Binder and Hidirolou (1988).

Gurney and Daly (1965) also generalized the results of Patterson (1950) to a linear model framework. They introduced the concept of an "elementary estimate". This is an estimate which uses data from a specific time period, based on individuals which all join and leave the survey at the same time. The expected value of these elementary estimates can be expressed as a linear combination of the population parameters, $\{\theta_t\}$. When the correlation structure is known, standard general linear model theory can be used to derive the MVLUE.

To formalize this discussion, let y_{tj} be the j -th elementary estimate from the t -th time period, where $E(y_{tj}) = \theta_t$. If Y and Θ are vectors with components y_{tj} and θ_t respectively, we can write:

$$Y = X'\Theta + e, \quad (2.1)$$

where X is a fixed $(n \times T)$ matrix of 0's and 1's, $E(e) = 0$ and $E(ee') = U$, which is the known variance-covariance matrix of the elementary estimates. Thus, the MVLUE is given by:

$$\tilde{\Theta} = (X'U^{-1}X)^{-1} X'U^{-1}Y, \quad (2.2a)$$

with

$$\text{Var}(\tilde{\Theta}) = (X'U^{-1}X)^{-1}. \quad (2.2b)$$

These results imply that every new survey would require the updating of all previous estimates. However, since estimates from the earlier occasions often have a much smaller effect than the recent occasions, composite estimates, such as proposed by Gurney and Daly (1965), are simpler to use and have a high relative efficiency. Binder and Hidirolou (1988) discussed the appropriateness of these methods and their application in a number of surveys. In general, they found that good results can be achieved using composite estimators, providing the rotation group biases are not substantial.

3. SIGNAL-NOISE EXTRACTION

It is quite common for economists and sociologists to treat the underlying parameters, $\{\theta_t\}$, as random inputs for their stochastic models (Smith 1978). However, if the sampling errors associated with the input data are ignored, the estimates of the parameters of the stochastic model are biased.

In this section, we show how the stochastic model assumptions can also be used to obtain model-dependent, design-consistent estimators. In Section 4, we discuss maximum likelihood estimation of these parameters. Since misspecification of the model could lead to serious biases,

hypothesis testing methods should be used to check the consistency of the model with the data. The model should also reflect the subject matter knowledge of the underlying phenomenon.

First we consider the case where the survey errors are independent. (This would be approximately true for non-overlapping surveys with small sampling fractions.) In this case, the MVLUE for θ_t is $\hat{\theta}_t = y_t$. However, by imposing a stochastic model for the sequence of parameters, $\{\theta_t\}$, an improvement in the mean squared error of the estimate can be achieved.

Scott and Smith (1974) proposed the following model for non-overlapping surveys. They wrote the model for the survey estimates at time t as:

$$y_t = \theta_t + e_t \quad (3.1)$$

where the e_t 's are independent $N(0, S_t^2)$. They assumed that the sequence of parameters, $\{\theta_t\}$, can be modelled such that, conditional on $\Theta'_{t-1} = (\theta_1, \dots, \theta_{t-1})$,

$$\theta_t = \underline{\alpha}_t' \Theta_{t-1} + \epsilon_t, \quad (3.2)$$

where the ϵ_t 's are independent $N(0, S_t^2)$ and independent of $\{e_t\}$, and $\underline{\alpha}_t$ is a $(t-1)$ dimensional vector of constants.

In general at time $t-1$, conditional on $Y'_{t-1} = (y_1, \dots, y_{t-1})$, we have $\Theta_{t-1} \sim N(\tilde{\Theta}_{t-1}, \tilde{V}_{t-1})$. Conditional arguments then yield

$$E(\theta_t | y_t) = \tilde{\theta}_t = \pi_t (\underline{\alpha}_t' \tilde{\Theta}_{t-1}) + (1 - \pi_t) y_t \quad (3.3a)$$

and

$$\text{Var}(\theta_t | y_t) = (1 - \pi_t) S_t^2, \quad (3.3b)$$

where

$$\pi_t = \frac{\text{Var}(y_t | \theta_t)}{\text{Var}(y_t)} = \frac{S_t^2}{\underline{\alpha}_t' \tilde{V}_{t-1} \underline{\alpha}_t + \sigma_t^2 + S_t^2}. \quad (3.3c)$$

Note that the estimator in (3.3a) is a weighted average of two components. The first consists of the best linear forecast of θ_t given the previous value of $\tilde{\Theta}_{t-1}$; the second consists of the best estimate of θ_t from the survey. The contribution of each term is controlled by π_t , the ratio of the survey variance to the total variance. As the survey error component becomes small, then the contribution from $\tilde{\Theta}_{t-1}$ becomes small and the estimate of θ_t in (3.3a) is composed primarily of y_t , the estimate from the survey data. Therefore, the estimator of θ_t is design-consistent whenever y_t is design-consistent.

However, as the survey error component becomes large, the estimate of θ_t is due primarily from the linear forecast of Θ_{t-1} . The relative efficiency of the estimator, $\tilde{\theta}_t$, in (3.3a) is given by $1/(1 - \pi_t)$, where π_t is defined in (3.3c). The greatest efficiency gains occur when the survey error is large relative to σ_t^2 , the variance of the "shocks" of the model process.

Scott and Smith (1974) and R.G. Jones (1980) also considered the case of overlapping surveys. Jones' formulation for this case was as follows. Let Θ_t be multivariate normal with mean zero and variance matrix V_t^* . Now the observations at time t may be generalized to a vector of elementary estimates, y_t . The conditional distribution of $Y_t = (y_1', \dots, y_t')'$ given Θ_t is assumed to be of the form:

$$Y_t = X_t' \Theta_t + e_t, \tag{3.4}$$

where X_t is a fixed matrix of 0's and 1's linking the parameters and the observations, and e_t is the survey error, assumed to be multivariate normal with mean zero and covariance matrix U_t .

Using conditional arguments, the best estimate of θ_t given Y_t is:

$$E(\Theta_t | Y_t) = \tilde{\Theta}_t = (X_t' U_t^{-1} X_t + \tilde{V}_t^{*-1})^{-1} X_t' U_t^{-1} Y_t \tag{3.5a}$$

with a variance of

$$\text{Var}(\Theta_t | Y_t) = (X_t' U_t^{-1} X_t + V_t^{*-1})^{-1}. \tag{3.5b}$$

This result is very general. If we allow the underlying stochastic model for Θ_t to be very diffuse, then the inverse of V_t^* is approximately zero, thus yielding the MVLUE given by (2.2a). R.G. Jones (1980) derived (3.5) by application of stochastic least squares, so that the estimator $\tilde{\Theta}_t$ is the minimum mean squared error (MMSE) linear estimator, even when the normality assumptions are dropped.

Applying (3.5) directly would involve inverting matrices which have the same dimensionality as the vector of all the elementary estimates for all time periods. Computing such inverses can be numerically unstable. However, expression (3.5) can often be restructured using state-space models, which are useful for describing many time series models. See Harvey (1984) for a review of such models. As we demonstrate below, this would avoid the inversion of large matrices. Some structure for $\{\theta_t\}$ and $\{e_t\}$ would be required to take advantage of the reduction in dimensionality afforded by the state-space approach. An example of such a structure, which is often used in time series applications, is an autoregressive-moving average (ARMA) process, not necessarily homogeneous in time.

For applications such as small area estimation, where the sample size is not large, modelling the variances of the survey error, U_t , using such ARMA models can be useful. This is not usually done for repeated surveys. This would also alleviate the problem of applying the result in (3.5) directly when the dimensions of V_t^* and U_t are large and the inverses are numerically unstable.

In the state-space model, two processes occur simultaneously. The first process, the observation system, details how the observations depend on the current state of the process parameters. The second process, the transition system, details how the parameters evolve over time.

State-space models can be written as follows. The observation equation is written as:

$$y_t = H_t z_t + \omega_t, \tag{3.6a}$$

and the transition equation is written as:

$$z_t = F_{t-1} z_t + G_t \underline{\epsilon}_t, \quad (3.6b)$$

where z_t is an $(r \times 1)$ state vector, H_t is a fixed $(n_t \times r)$ matrix, F_t is a fixed $(r \times r)$ transition matrix, G_t is a fixed $(r \times m)$ matrix and $\underline{\omega}_t$ and $\underline{\epsilon}_t$ are independent random disturbances with mean zero and covariances given by $E(\underline{\omega}_t \underline{\omega}_t') = U_t$ and $E(\underline{\epsilon}_t \underline{\epsilon}_t') = V_t$.

As an example of this formulation, we rewrite the model studied by Blight and Scott (1973) in terms of the state-space model. Blight and Scott considered data from Patterson's (1950) one-level rotation design. They let \bar{y}_t'' be the mean of the new units at time t , and \bar{y}' and \bar{x}_t' the means of the overlapping units at times t and $t-1$, respectively. They assumed that \bar{y}_t'' and $\bar{y}_t' - \rho \bar{x}_t'$ are independent observations at time t , where ρ is the between-occasion correlation of the responses from the same individual. They also assumed that the mean process $\{\theta_t\}$ is first order autoregressive.

We let the state vector be $z_t' = (\theta_t, \theta_{t-1})$. The observation equation can be written as:

$$\begin{bmatrix} \hat{y}_t'' \\ \bar{y}_t' - \rho \bar{x}_t' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\rho \end{bmatrix} \begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} + \begin{bmatrix} \omega_{1t} \\ \omega_{2t} \end{bmatrix},$$

where $(\omega_{1t}, \omega_{2t})'$ has a diagonal covariance matrix.

The transition equation would be written as:

$$\begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_{t-1} \\ \theta_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t,$$

where ϵ_t is $N(0, \sigma^2)$. Thus, the Blight-Scott model can be written in state-space form.

Harvey and Phillips (1979) described a method to put the ARMA (p, q) model, defined by:

$$y_t - \alpha_1 y_{t-1} - \dots - \alpha_p y_{t-p} = \epsilon_t - \beta_1 \epsilon_{t-1} - \dots - \beta_q \epsilon_{t-q}, \quad (3.7)$$

where the ϵ_t 's are independent $N(0, \sigma^2)$, into state-space form. The dimension of z_t is $r = \text{MAX}(p, q+1)$. Where necessary, $\underline{\alpha} = (\alpha_1, \dots, \alpha_p)$ or $\underline{\beta} = (\beta_1, \dots, \beta_q)$ is augmented with zeroes to have dimension r . The matrix, U_t is set to zero. The ARMA (p, q) model is equivalent to (3.6) when $H_t' = (1, 0, \dots, 0)$, $G_t' = (1, -\beta_1, \dots, -\beta_{r-1})$ and

$$F_t = \left[\begin{array}{c|c} \alpha_1 & \\ \vdots & I_{r-1} \\ \alpha_{r-1} & \\ \hline \alpha_r & O' \end{array} \right],$$

where I_{r-1} the $(r-1) \times (r-1)$ identity matrix and O' is a row vector of zeroes.

In this formulation, the state vector $z_t = (z_{1t}, \dots, z_{rt})'$ is defined as follows:

$$z_{it} = \alpha_i y_{t-1} + \alpha_{i+1} y_{t-2} + \dots + \alpha_r y_{t-(r-i+1)} \\ - \beta_{i-1} \epsilon_t - \beta_i \epsilon_{t-1} - \dots - \beta_{r-1} \epsilon_{t-(r-i)} ,$$

for $i = 2, 3, \dots, r$ and $z_{1t} = y_t$ as in (3.7).

A necessary condition for stationarity is that $\text{Var}(z_t) = \text{Var}(z_{t-1})$ for all t . From expression (3.6b), we see that this implies that

$$\text{Var}(z) = F' \text{Var}(z) F + G V G' ,$$

where $V_t \equiv V$ is constant for all t . Pearlman (1980) pointed out that this can be used to obtain the initial conditions for z_1 .

Often the survey error process can be included in the state-space model, when some structure for the survey errors can be assumed. We have already demonstrated this for the Blight and Scott (1973) model. Scott and Smith (1974) and Miazaki (1985) considered a variety of models which were special cases of $\{\theta_t\}$ being ARMA (p, q) , $\{e_t\}$ being ARMA (p^*, q^*) and the scalar observations satisfying $y_t = \theta_t + e_t$. State-space models for this process can be formulated analogously to the Harvey-Phillips representation above, where the state vector z_t is the vector formed by concatenating the state vectors from each of the individual ARMA processes.

For example, suppose $\{\theta_t\}$ is an ARMA $(3, 0)$ process with parameter $(\alpha_1, \alpha_2, \alpha_3)$ and model variance σ^2 and, $\{e_t\}$ is an ARMA $(0, 1)$ process with parameter β^* and model variance s^2 . An ARMA $(0, 1)$ process for $\{e_t\}$ would be plausible for a survey which follows Eckler's two-level rotation sampling pattern, where the survey estimate for θ_t is given by \bar{y}_t , the mean of all individuals reporting for the t -th occasion.

This can be written in state-space form by letting

$$F_t = \left[\begin{array}{ccc|ccc} \alpha_1 & 1 & 0 & 0 & 0 & 0 \\ \alpha_2 & 0 & 1 & 0 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right], G_t = \left[\begin{array}{c|c} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ \hline 0 & 0 \\ 0 & -\beta^* \end{array} \right], V_t = \left[\begin{array}{c|c} \sigma^2 & 0 \\ \hline 0 & s^2 \end{array} \right], \quad (3.8)$$

$U_t = 0$ and $H_t' = (1 \ 0 \ 0 | 1 \ 0)$. The first three components of the state vector correspond to the state-space formulation for the $\{\theta_t\}$ process and the last two components are for the $\{e_t\}$ process.

Note that the state-space approach allows for measurement error, given by ω_t in (3.6a). However, unless the survey design has non-overlapping units with independent sampling errors, the measurement error terms cannot be used to model the survey error. Instead, we have absorbed the measurement (survey) error into the state vector.

From the general state-space framework, the Kalman filter equations can be derived. If, as in Meinhold and Singpurwalla (1983), we let the conditional distribution of z_{t-1} given Y_{t-1} be $N(\tilde{z}_{t-1|t-1}, \tilde{P}_{t-1|t-1})$, then recursive relationships for $\tilde{z}_{t|t}$ and $\tilde{P}_{t|t}$ can be constructed. Harvey (1984) shows these relationships are equivalent to the Kalman filter.

The Kalman filter, in general, consists of two parts. The first is a one-step ahead prediction of the state vector and its covariance; the second part provides an update of the mean and covariance matrix of the state-space vector after the new observations are available.

Following the notation used in (3.6), we let $Y_1 = y_1$ and $Y'_{t+1} = (Y'_t, y'_{t+1})'$, then the one-step ahead prediction has a mean and variance given by

$$E(z_1) = \tilde{z}_{1|0} \quad (3.9a)$$

$$\text{Var}(z_1) = \tilde{P}_{1|0} \quad (3.9b)$$

$$E(z_t | Y_{t-1}) = \tilde{z}_{t|t-1} = F_t \tilde{z}_{t-1|t-1} \quad (3.9c)$$

$$\text{Var}(z_t | Y_{t-1}) = \tilde{P}_{t|t-1} = F_t \tilde{P}_{t-1|t-1} F'_t + G_t V_t G'_t. \quad (3.9d)$$

The update of the mean and variance for the state vector at time t after the observation at time t becomes available is:

$$\begin{aligned} E(z_t | Y_t) &= \tilde{z}_{t|t} \\ &= \tilde{z}_{t|t-1} + \tilde{P}_{t|t-1} H'_t (H'_t \tilde{P}_{t|t-1} H_t + U_t)^{-1} (y_t - H'_t \tilde{z}_{t|t-1}) \end{aligned} \quad (3.10a)$$

$$\text{Var}(z_t | Y_t) = \tilde{P}_{t|t} = \tilde{P}_{t|t-1} - \tilde{P}_{t|t-1} H_t (H'_t \tilde{P}_{t|t-1} H_t + U_t)^{-1} H'_t \tilde{P}_{t|t-1} \quad (3.10b)$$

The equations (3.9) and (3.10) are the well-known Kalman filter equations. The formulation followed here is essentially Bayesian; however, it is possible to derive equivalent results using orthogonal projections; see Young (1984).

The simplification in the computations due to the Kalman filter formulation in the sample survey setting can be seen by comparing equations (3.9) and (3.10) with R.G. Jones' (1980) result (3.5). Note that Jones' result required the inversion of a matrix with dimensionality given by the complete vector of survey estimates.

The Kalman filter can also be used to obtain smoothed estimates given by $E(z_t | Y_t)$ for $T > t$. Details of this backcasting may be found in Harvey (1984).

Remarks

1. Although the Kalman filter assumes an infinite population model, when the sample survey is based on a large sample, the central limit theorem often allows the survey errors to be approximately normally distributed. As well, since the smoothed estimators for $\{\theta_t\}$ are the same as those obtained by R.G. Jones (1980) in (3.5a), these are the linear MMSE estimators even when the normality assumptions are dropped.

2. Missing time points can be incorporated in the state-space approach. If y_t is missing at time t , then the updating equations analagous to (3.9) become $\tilde{z}_{t|t} = \tilde{z}_{t|t-1}$ and $\tilde{P}_{t|t} = \tilde{P}_{t|t-1}$ as in R.H. Jones (1980). However, smoothed estimates for the missing time points will depend strongly on the model selected, since no survey estimate is available. Therefore, the risks of model misspecification here are high.
3. The likelihood function, which we discuss in Section 4 for obtaining the maximum likelihood estimates of the unknown parameters, can also be obtained when some data are missing, using the same approach given by R.H. Jones (1980). However, missing data will tend to increase the standard errors of the parameter estimates. In our example of Section 5, we encounter a case with missing time points.

4. ESTIMATION OF THE PARAMETERS IN A STATE-SPACE MODEL

When data are generated from the ARMA model (3.7) and the parameters α , β , and σ^2 are unknown, the maximum likelihood estimates for the unknown parameters can be obtained using the likelihood function derived from the state-space model. This approach was suggested by Harvey and Phillips (1979), R.H. Jones (1980) and others.

The usual state-space models can also be used when the input data have independent measurement errors. This is the case for our example of Section 5, where we show the effect on the parameter estimates when the survey errors are taken into account.

Maximum likelihood estimation of these parameters when the data have correlated survey errors has not previously been studied in detail. For a model with univariate stationary observations $\{y_t\}$, Scott, Smith and Jones (1977) suggested using the estimated autocovariance function of the observations $\{y_t\}$ to estimate the parameters of the ARMA process. Here, the data model is $y_t = \theta_t + e_t$. The variances and covariances of the survey errors, $\{e_t\}$ can be estimated using design-based methods; see, for example, Wolter (1985).

Efficient estimation of the autocovariances of the survey errors, assuming stationarity of the series, is an area which has not received attention in the literature, so ad hoc methods would be used in practice. Future research in modelling these survey errors would be worthwhile. In our example in Section 5, we could assume independent survey errors, so this was not problematic.

Assuming the autocovariance of $\{e_t\}$ is available, the autocovariance of $\{\theta_t\}$ can be estimated by $\text{Cov}(\theta_t, \theta_{t-s}) = \text{Cov}(y_t, y_{t-s}) - \text{Cov}(e_t, e_{t-s})$. However, this method is not fully efficient (Smith; 1978). Moreover, this method would not incorporate non-stationary survey errors.

Miazaki(1985) considered the case where $\{\theta_t\}$ is an ARMA $(p,0)$ process. She also assumed $\{e_t\}$ to be an ARMA $(0,q)$ process which could be estimated directly from the survey. Miazaki then wrote the observations $\{y_t\}$ as an ARMA $(p,p+q)$ process which she estimated by restricted maximum likelihood methods.

Representing non-stationarity of survey errors in the state-space representation can sometimes be handled through nonhomogeneous matrices for V_t , the variance matrix of the random "shocks" from the transition equation (3.6b). For example, in (3.7) s^2 would be replaced by s_t^2 to allow for non-homogeneous survey errors. This approach is taken in the example in Section 5.

In general, for state-space models given by (3.5), Harvey and Phillips (1979) write the exact likelihood function as follows. Letting

$$\hat{y}_{t|t-1} = E(y_t | Y_{t-1}) = H'_t \tilde{z}_{t|t-1}$$

and

$$R_t = \text{Var}(y_t | Y_{t-1}) = H'_t \tilde{P}_{t|t-1} H_t + U_t,$$

the log-likelihood function for $Y'_T = (y'_1, \dots, y'_T)$ is

$$\log f(Y_T) = (1/2) \sum_{t=1}^T \log |R_t| - (1/2) \sum_{t=1}^T (y_t - \hat{y}_{t|t-1})' R_t^{-1} (y_t - \hat{y}_{t|t-1}). \quad (4.1)$$

The unknown parameters in (4.1) are contained in $\hat{y}_{t|t-1}$ and in R_t . Depending on the algorithm used to maximize (4.1) with respect to the unknown parameters, it may be necessary to compute first and second derivatives of (4.1) with respect to the unknown parameters. This generally involves finding derivatives of $\tilde{z}_{t|t-1}$ and $\tilde{P}_{t|t-1}$. These can be computed numerically using the recursions given in (3.8) and (3.9). For example, (3.8c) yields $\partial \tilde{z}_{t|t-1} = (\partial F_t) \tilde{z}_{t-1|t-1} + F_t (\partial \tilde{z}_{t-1|t-1})$. The other expressions using (3.8) and (3.9) can be determined similarly.

The inclusion of regression parameters into (4.1) can be accomplished by replacing y_t by the deviation of y_t from the regression line. Tam (1987) generalized this concept even further by considering a model where the underlying stochastic process is determined by a state-space model for the regression coefficients which evolve over time.

To maximize the likelihood function (4.1) with respect to unknown parameters, an iterative procedure is needed. We omit details of the procedure used for the application in Section 5 since efficient procedures are still in the development stage.

Once having estimated the parameters, smoothed values for the state vector, $\tilde{z}_{t|T} = E(z_t | Y_T)$ after time $T > t$, can be obtained using the backcasting formulae given by the Kalman filter; see Harvey (1984). Thus, for example, if $y_t = \theta + e_t$ as in (3.1), after backcasting we may formulate $y_t = \tilde{\theta}_{t|T} + \tilde{e}_{t|T}$, so that $\tilde{\theta}_{t|T}$ becomes the smoothed estimate of the mean at time t after observing Y_T .

To derive the standard error of the smoothed estimate it is necessary to account for the fact that the unknown parameters have been estimated from the data, particularly when the data series is short; see Jones (1979). Hamilton (1986) suggests doing this by Monte Carlo simulations. He generates a set of multivariate normal random variables with mean given by the maximum likelihood estimates for the parameters and variance given by the inverse of the estimated Fisher information matrix. He then estimates $E(\tilde{P}_{t|T})$ and $\text{Var}(\tilde{z}_{t|T})$, where the expectation and variance are taken over the generated parameter values. The sum of these two components is the estimated covariance matrix of the estimated state vector. This method assumes that the sample size is large, so that the normal approximation to the sampling distribution of the parameter estimates is valid.

In the examples of Section 5, we approximate the standard deviation of the sampling errors of the smoothed estimates, ignoring the variation due to estimating certain model parameters. We then compare these with the actual root mean squared errors of the sampling distribution obtained from simulated data.

5. DATA ANALYSIS

In this section we show the impact of the survey errors on estimates of the parameters of a first order autoregressive model with regression terms. In our example the survey errors are assumed to be independent between occasions. More complicated cases with correlated survey error and higher order ARMA models for the population characteristic could be handled within the framework we have described. We chose this example to demonstrate that the impact of accounting for the survey errors can be appreciable even for this relatively simple model.

We used data from Saskatchewan respondents to the Canadian Travel Survey (CTS). The CTS is conducted by Statistics Canada to collect descriptive statistics on the travelling habits and characteristics of Canadian residents. This survey is conducted as an "add-on" to the Labour Force Survey (LFS). The LFS is a monthly rotating panel survey with six rotation groups. However, the CTS is conducted at most four times a year, with at least one, but possibly as many as three rotation groups. The rotation groups used by the CTS for the quarters when the CTS is conducted are chosen so that there are no overlapping panels between occasions.

The survey errors are assumed to be independent. This is only approximately true. The LFS is a multi-stage survey and the primary sampling units (PSU's) do not rotate out as quickly as the individual rotating panels. The same PSU's are used on a number of occasions. Therefore, although the CTS sample is selected such that the panels do not overlap between occasions, the independence assumption is approximately true only when the correlation of the sampling errors between quarterly periods within the same PSU is small. This assumption was not verified.

The coefficients of variation (as a percentage) were calculated using the function:

$$CV = \alpha y^{-\beta} / \sqrt{\text{number of rotation groups}},$$

where y is the survey estimate in thousands. This is the function recommended to users of the CTS for data on Saskatchewan residents; see Statistics Canada (1985). In this report, the parameters α and β were estimated at 91.7528 and 0.353253, respectively, using a loglinear regression model applied to 1979 data. For the purposes of our example, these coefficients of variation were rounded to the nearest tenth of a percent.

The assumed model was:

$$y_t = \theta_t + e_t, \quad (5.1)$$

where the e_t 's are independent survey errors, with $e_t \sim N(0, s_t^2)$ and

$$\theta_t = \gamma_0 + \gamma_1 t + \gamma_2 Q_{1t} + \gamma_3 Q_{2t} + \gamma_4 Q_{3t} + \epsilon_t, \quad (5.2)$$

where $\{\epsilon_t\}$ is ARMA (1,0) with parameters (α, σ^2) . The regression terms in (5.2) are, respectively, the intercept, a term representing the quarter number with t taking values from -15.5 to 15.5 linearly in time and, finally, seasonal terms for the first three quarters of each year, where

$$\begin{aligned} Q_{it} &= 1 \text{ if the } t\text{-th observation is in the } i\text{-th quarter;} \\ &= -1 \text{ if the } t\text{-th observation is in the fourth quarter;} \\ &= 0 \text{ otherwise;} \end{aligned}$$

for $i = 1, 2, 3$.

Better models may be available for these data, although with such a small data set, tests of hypotheses against alternative models would not be very powerful.

To obtain the maximum likelihood estimates for the unknown parameters of this model, it is necessary to incorporate the assumptions made about the survey errors in the estimation procedure. Most users of official statistics ignore this survey error and implicitly assume that the input data are error-free. This does not seriously affect the results when the variance of the survey error is small relative to the variance of the model error.

The survey estimates and the coefficients of variation of the survey errors relative to these estimates are given in Tables 1 and 2. The results of the maximum likelihood estimation procedure are displayed in Tables 3 and 4. Two estimates are given for each model. The column labeled "Estimate: With Sampling Error" uses the method incorporating the assumed error structure; whereas the column labeled "Estimate: Ignoring Sampling Error" repeats the estimation under the assumption that the survey estimate is observed without error. In both cases model (5.2) is assumed.

Table 1
Overnight Person-Trips of Saskatchewan Residents to
Destinations within Saskatchewan¹

Year	Quarter	No. of Rotation Groups	Survey Estimate (000's)	Smoothed Estimate (000's)	Survey C.V. (%)	Smoothed C.V. (%)	Simulated RMSE (%)	Simulated Bias (%)
1979	Winter	1	598	611	9.6	5.9	6.9	0.1
	Spring	1	808	813	8.6	4.8	4.9	0.4
	Summer	3	1033	1103	4.6	3.0	3.1	0.0
	Fall	3	678	683	5.3	4.3	4.5	1.2
1980	Winter	1	578	608	9.7	5.5	5.8	0.1
	Spring	3	837	837	4.9	3.7	3.6	0.0
	Summer	1	1451	1169	7.0	3.3	3.5	0.3
	Fall	1	744	724	8.9	5.1	5.9	0.8
1981	Winter	3	631	632	5.4	4.3	5.0	-0.1
	Summer	3	1262	1172	4.2	2.9	3.3	0.1
1982	Winter	1	565	613	9.8	5.5	6.4	-0.4
	Spring	1	901	838	8.3	4.5	5.1	0.8
	Summer	3	1167	1147	4.4	2.9	3.1	0.1
	Fall	1	721	706	9.0	5.1	5.6	0.2
1984	Winter	1	585	598	9.6	5.8	6.7	-1.2
	Spring	1	788	804	8.7	4.6	5.2	-0.4
	Summer	3	1068	1107	4.5	2.9	3.6	-0.5
	Fall	1	711	686	9.0	5.3	6.7	0.7
1986	Winter	1	793	630	8.7	6.2	7.1	-1.3
	Spring	3	798	808	5.0	3.9	3.9	-0.4
	Summer	3	1053	1096	4.5	3.0	3.3	-0.3
	Fall	3	650	663	5.4	4.4	4.2	0.2

¹ The Canadian Travel Survey was not conducted in the Spring and Fall Quarters of 1981 and during 1983 and 1985.

Simulations in last two columns are based on a sample size of 100.

Table 2
Overnight Person-Trips of Saskatchewan Residents to
Destinations in Manitoba¹

Year	Quarter	No. of Rotation Groups	Survey Estimate (000's)	Smoothed Estimate (000's)	Survey C. V. (%)	Smoothed C. V. (%)	Simulated RMSE (%)	Simulated Bias (%)
1979	Winter	1	27	34	28.6	13.4	14.1	0.5
	Spring	1	33	48	26.7	11.0	10.2	0.9
	Summer	3	78	80	11.4	6.6	7.1	1.3
	Fall	3	55	48	12.9	10.1	10.8	0.6
1980	Winter	1	24	30	29.7	13.6	14.5	0.5
	Spring	3	63	50	12.3	9.5	9.4	0.7
	Summer	1	86	80	19.0	6.6	6.3	0.8
	Fall	1	75	46	19.9	11.0	12.2	0.5
1981	Winter	3	42	34	14.2	11.3	13.2	1.0
	Summer	3	79	82	11.3	5.9	5.7	0.1
1982	Winter	1	33	34	26.5	12.5	13.2	-2.8
	Spring	1	46	44	23.7	10.7	10.0	1.6
	Summer	3	78	82	11.4	5.7	5.4	0.1
	Fall	1	30	42	27.6	10.9	11.4	0.3
1984	Winter	1	36	34	25.7	13.8	16.8	-1.3
	Spring	1	48	43	23.4	11.4	11.5	0.1
	Summer	3	82	82	11.1	6.1	7.3	-0.2
	Fall	1	30	40	27.7	11.5	11.4	0.6
1986	Winter	1	33	33	26.7	16.3	19.9	-0.8
	Spring	3	38	41	14.6	10.9	11.7	-0.1
	Summer	3	90	81	10.8	7.1	8.8	-0.3
	Fall	3	42	40	14.1	11.2	10.5	1.7

¹ The Canadian Travel Survey was not conducted in the Spring and Fall Quarters of 1981 and during 1983 and 1985. Simulations in last two columns are based on a sample size of 100.

Table 3
Parameter Estimates for Saskatchewan to Saskatchewan Person-Trips¹

Parameter	Ignoring Sampling Error	With Sampling Error				
	Estimate	Estimate	Standard Error	Simulated RMSE	Simulated Bias	t-value of Bias
REGRESSION						
Intercept (γ_0)	831.4	815.0	15.6	14.4	1.8	1.29
Linear (γ_1)	-0.84	-0.86	1.52	1.51	-0.10	-0.65
1st Quarter (γ_2)	-209.6	-203.8	21.8	24.6	-3.5	-1.41
2nd Quarter (γ_3)	-4.0	7.1	22.9	23.8	0.4	0.17
3rd Quarter (γ_4)	340.1	316.0	21.2	23.4	-0.4	-0.18
ARMA						
Autoregressive (α)	0.14	0.47	0.66	0.68	-0.39	-6.77
Model Variance (σ^2)	7930.5	879.3	1205.6	770.0	-488.2	-8.16

¹ Simulations and t-values are based on a sample size of 100.

Table 4
Parameter Estimates for Saskatchewan to Manitoba Person-Trips¹

Parameter	Ignoring Sampling Error	With Sampling Error				
	Estimate	Estimate	Standard Error	Simulated RMSE	Simulated Bias	<i>t</i> -value of Bias
REGRESSION						
Intercept (γ_0)	51.2	50.5	1.9	2.0	0.4	1.57
Linear (γ_1)	-0.17	-0.13	0.18	0.17	-0.04	-2.01
1st Quarter (γ_2)	-20.1	-17.2	3.4	3.5	-0.6	-1.52
2nd Quarter (γ_3)	-5.9	-6.1	3.6	3.7	-0.1	-0.32
3rd Quarter (γ_4)	30.7	30.8	3.7	3.7	0.0	-0.07
ARMA						
Autoregressive (α)	0.14	-0.75	0.66	0.71	0.49	7.90
Model Variance (σ^2)	100.0	5.7	18.7	9.5	-0.3	-0.29

¹ Simulations and *t*-values are based on a sample size of 100.

The estimates of the regression parameters are essentially the same under either assumption. However, the autoregressive component estimates differ considerably under the two assumptions. In particular, the model variance increases substantially. This variance estimate increases because the variation due to survey error is missing from the model. The reason that the estimates of the regression coefficients are not affected is that the estimators for these coefficients remain unbiased, although they are somewhat inefficient.

Once the parameters of the model have been estimated, it is possible to use the assumed model to adjust the individual estimates of the number of overnight person-trips. The results discussed below demonstrate how the procedure reduces the coefficients of variation for these smoothed estimates when the model assumptions are correct. Such a procedure is analogous to model-dependent small area estimation methods.

The smoothed estimates and their coefficients of variation are given in Tables 1 and 2. These coefficients of variation are calculated, taking into account the sampling error of the regression coefficients, $\gamma_0, \dots, \gamma_4$. This is possible since, given α and σ^2 , the smoothed estimates are linear functions of the original survey estimates, so that the variances can be computed from this linear function and the assumed model variance of the regression residuals. However, the sampling errors for the estimated α and σ^2 were ignored at this point. The effect of ignoring these sampling errors is discussed below.

The smoothed estimates for travel within Saskatchewan are generally close to the original survey estimates, with possible exceptions for the Summer of 1980 and the Winter of 1986. Those for travel to Manitoba are also close, with a possible exception being the Fall of 1980. These exceptional cases could possibly be outliers or could be due to a special event that boosted tourism in those quarters. In general, such phenomena could be incorporated into the model by: (i) increasing the model variance in the state-space model for those periods or adding appropriate dummy variables for special events or (ii) increasing the sampling variance for outliers. A more in-depth knowledge of the circumstances would be required to decide whether such adjustments are appropriate. The analysis here can help pinpoint possible unusual cases.

Because the analysis so far has ignored the effect of the sampling error associated with estimating α and σ^2 , we performed a simulation study to assess its seriousness. Jones (1979), Hamilton (1986) and Tam (1987) have suggested that these sampling errors should not be ignored, especially when the time series has few observations.

For the simulation, we generated sets of random data following the assumed model given by (5.1) and (5.2). We took as our parameter values the maximum likelihood estimates of the model. The same missing data pattern was used in the simulations as in the original data set. One hundred such data sets were generated for each model. In Tables 1 and 2, we report the percentage bias of the smoothed values and the percentage root mean squared error for the difference between the smoothed values and the true values based on these simulations.

To assess whether 100 was a sufficiently large number of simulations to estimate the root mean squared error (RMSE), we computed an estimate of the coefficient variation of the estimator of the RMSE. From the simulations we obtained an unbiased estimate of the variance of the estimator of the mean squared error. We then used Taylor linearization to estimate the variance of the estimator of the RMSE. The estimated coefficients of variation ranged from 6% to 11% for destinations within Saskatchewan and from 5% to 9% for destinations in Manitoba. Therefore, these estimates of the RMSE's do provide a reasonable assessment of the effect of ignoring the sampling error of the autoregressive parameters.

In Tables 1 and 2, the biases of the adjustment procedure are all small and, in fact, for the two sets of 22 observations only four were significant at the 5% level using a standard *t*-test.

We also note that the percentage root mean squared errors based on the imulations tend to be larger than those under the column entitled "Smoothed C.V.". This is to be expected since the simulations include sampling errors arising from the estimation of α and θ^2 . However, the values of the "Smoothed C.V.'s" do give reasonable approximations to the simulated values, so the procedure which ignores the effect of the sampling error of α and θ^2 does not seriously affect the coefficients of variation.

In Table 3 and 4, we report some simulation results for the estimated parameters. For the regression coefficients, only one of the biases was significant at the 5% level. The standard errors are all consistent with the simulation results.

On the other hand, the simulations did point out a problem with the estimates for α and σ^2 . The biases for the estimates of α were highly significant. As can be seen from Tables 3 and 4, one of the biases of σ^2 was also highly significant. The simulated root mean squared errors were not very close to the asymptotic approximation of the standard error obtained by inverting the Fisher information matrix. It seems that the sample size for our problem is not sufficiently large for the asymptotic approximations to be very accurate. This is a common problem for time series analyses of short series.

6. CONCLUSION

In cases where the variances of the survey errors are small relative to the variances of the model errors, the smoothed estimates would be close to the minimum variance linear unbiased estimates and there would be no appreciable reduction in the standard errors of the estimates, even when the assumed model is true. However, for cases such as small domain estimation where the sampling errors are not small, the standard errors for the smoothed estimates may be substantially smaller than those for the original survey estimates. For example, the smoothed estimates for the Saskatchewan-to-Manitoba data showed a greater improvement than the Saskatchewan-to-Saskatchewan data, since the sampling errors for the survey data were larger for the former data set.

One of the implications of assuming models for repeated surveys is that if the models are misspecified, the MMSE estimators may be seriously biased. It is important, therefore, to choose a model which is both consistent with the data and which reflects subject matter knowledge about the underlying phenomena. In our example the data set is small, so that a large number of statistical models would be consistent with the data.

Our simulation studies suggest that even for small data sets, the asymptotic approximations to the variances of the smoothed estimates are quite reasonable. However, as in the case of more traditional applications of time series analyses, the asymptotic approximations for the sampling errors of the parameter estimates may be poor.

ACKNOWLEDGEMENTS

The authors would like to thank an Associate Editor of this journal and the referees for helpful comments on earlier versions. In particular, we are grateful to one referee whose thorough and insightful comments have led to many improvements of the paper. We would also like to thank Pierre Hubert, Chief of the Tourism, Travel and Recreation Section, Education Culture and Tourism Division for making available the data from the Canadian Travel Survey. Some of the material presented here appeared in the second author's Master thesis at the University of Guelph.

REFERENCES

- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in Time. In *Handbook of Statistics, Vol. 6*, (Eds, P.R. Krishnaiah and C.R. Rao), Amsterdam: Elsevier Science, 187-211.
- BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Ser. B*, 35, 61-68.
- ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-685.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 247-257.
- HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33, 387-397.
- HARVEY, A.C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3, 245-275.
- HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.
- JONES, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389-395.
- KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82, 35-45.

- MEINHOLD, R.J., and SINGPURWALLA, N.D. (1983). Understanding the Kalman Filter. *The American Statistician*, 37, 123-127.
- MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. Ph.D. Thesis, Iowa State University, Ames, Iowa.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Ser. B*, 12, 241-255.
- PEARLMAN, J.G. (1980). An algorithm for the exact likelihood of a high-order autoregressive-moving average process. *Biometrika*, 67, 232-233.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F., and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistics Review*, 45, 13-28.
- SINGH, D. (1968). Estimates in successive sampling using multi-stage design. *Journal of the American Statistical Association*, 63, 99-112.
- SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. In *Survey Sampling and Measurement*, (Ed. N.K. Namboodini), New York: Academic Press, 201-216.
- STATISTICS CANADA (1985). Canadian Travel Survey: Estimation and variance estimation procedures. Technical Report, Statistics Canada.
- TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.
- TIKKIWAL, B.D. (1979). Successive sampling — a review. *Bulletin of the International Statistics Institute*, 48, 367-384.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- YOUNG, P. (1984). *Recursive Estimation and Time Series Analysis: An Introduction*. New York: Springer-Verlag.

Sample Allocation in Multivariate Surveys

JAMES BETHEL¹

ABSTRACT

The optimum allocation to strata for multipurpose surveys is often solved in practice by establishing linear variance constraints and then using convex programming to minimize the survey cost. Using the Kuhn-Tucker theorem, this paper gives an expression for the resulting optimum allocation in terms of Lagrangian multipliers. Using this representation, the partial derivative of the cost function with respect to the k -th variance constraint is found to be $-2\alpha_k^* g(x^*)/v_k$, where $g(x^*)$ is the cost of the optimum allocation and where α_k^* and v_k are, respectively, the k -th normalized Lagrangian multiplier and the upper bound on the precision of the k -th variable. Finally, a simple computing algorithm is presented and its convergence properties are discussed. The use of these results in sample design is demonstrated with data from a survey of commercial establishments.

KEY WORDS: Multiple objective sample allocation; Nonlinear programming; Stratified sampling.

1. INTRODUCTION

The problem of optimum sample allocation in surveys with multiple study objectives was first discussed by Neyman (1934) in his development of the theory for solving the univariate optimum allocation problem. Since then, many researchers have studied the multivariate problem and several approaches have been suggested, most of which fall into one of two categories. The first involves forming a weighted average of the stratum variances and finding the optimal allocation for the "average variance" which results. Dalenius (1953), Yates (1960), Folks and Antle (1965), Hartley (1965), and Kish (1976) discuss methods related to this approach. The second basic technique is to require that each variance satisfy an inequality constraint and then use convex programming to obtain the least cost allocation which satisfies all the constraints. Dalenius (1957), Yates (1960), Kokan (1963), Hartley (1965), Kokan and Khan (1967), Chatterjee (1968, 1972), Huddleston, Claypool, and Hocking (1970), Bethel (1985), and Chromy (1987) all discuss the use of convex programming in relation to the multivariate optimal allocation problem. Each approach has its advantages and disadvantages. The "weighted average" method is computationally simple, intuitively appealing, and can be solved under a fixed cost assumption, but the choice of the weights is arbitrary and the optimality properties are not clear. The "convex programming" approach gives the optimal solution to the defined problem but the resulting cost may not be acceptable so that a further search is usually required for an optimal solution which falls within the budgetary constraints.

In this paper, a closed expression for the optimal allocation subject to linear inequality constraints will be given in terms of Lagrangian multipliers. In this framework, two results easily follow which substantially overcome the disadvantages of the convex programming approach. The first is that scaling the optimal multivariate allocation results in an allocation which is optimal under constraints which are proportionate to the original ones. Thus, if the optimal solution is too costly, it can be scaled down to the allowable budget directly and the effects of this on the precision of sample estimates can be directly determined. The second result is

¹ James Bethel, Westat, 1650 Research Boulevard, Rockville, MD. 20850 USA.

a simple expression for the partial derivatives of the cost of the sampling allocation with respect to the variance constraints. These quantities, called "shadow prices", show the sensitivity of the cost to variance constraints and are useful in assessing the cost effectiveness of the sample design.

The problem of solving the convex optimization still remains. Much has been written on methods for solving programming problems of this type and there are many software packages available for doing so. Some special programming considerations will be discussed here, however, and a simple method will be presented. This algorithm, essentially a steepest descent procedure, is convergent, straightforward to program, and easy to use, since no initial values are required. An example will be presented which demonstrates this algorithm and the other techniques discussed above.

2. THE ALLOCATION MODEL

Consider the case of stratified random sampling with I strata and J variables. Suppose it is required that the j -th variable satisfy

$$\text{Var}(\bar{y}_j) \approx \sum_{i=1}^I W_i^2 S_{ij}^2 / n_i \leq v_j^2, \quad (1)$$

where S_{ij}^2 , n_i , and W_i^2 , are, respectively, the variance of the j -th response variable, the sample allocation, and the proportion of the population that fall in the i -th stratum, and where v_j is an arbitrary, positive constant. In this paper it will be assumed that the finite population correction factors are negligible. In practice, it is expected that the effects of this assumption, which will be discussed in more detail in Section 7, would be limited.

Let

$$\begin{aligned} x_i &= 1/n_i \text{ if } n_i \geq 1 \\ &= \infty \text{ otherwise} \end{aligned}$$

and assume the cost function

$$g(x) = \sum_{i=1}^I c_i / x_i, \quad c_i > 0, \quad i = 1, 2, \dots, I. \quad (2)$$

A constant term for fixed costs could be included, but this would not affect the minimization process and is deleted here to simplify the notation. Define the constants

$$a_{ij} = w_i^2 S_{ij}^2 / v_j^2 \quad (3)$$

which will be referred to as "standardized precision units". Notice that $a_{ij} \geq 0$. Using this notation, the optimal allocation problem can be expressed as follows:

$$\begin{aligned} &\text{Minimize} && g(x) \\ &\text{subject to} && a'_j x \leq 1, \quad j = 1, 2, \dots, J \\ & && x > 0 \end{aligned} \quad (4)$$

where a_j is the j -th column vector of the matrix $A = \{a_{ij}\}$.

Kokan (1963) discusses this allocation model extensively and shows how it can be adapted to cover many common sample allocation problems, including cluster sampling and double sampling. Kokan and Khan (1967) give further analytical results in this context; Arthanari and Dodge (1981) restate Kokan and Khan's results. In related work, Kish (1976) describes a class of "linear forms" which occur frequently in survey research and to which many of the results developed here will apply.

3. THE OPTIMUM ALLOCATION

The optimum allocation for a single variable is well known. In that case $J = 1$, and the minimum of $g(x)$ subject to $a'_j x \leq 1$ with $x > 0$, denoted by x^* , is given by

$$\begin{aligned} x_i^* &= \sqrt{c_i} / \left(\sqrt{a_{i1}} \sum_{k=1}^I \sqrt{c_k a_{k1}} \right) && \text{if } a_{i1} > 0, 1 \leq i \leq I \\ &= \infty && \text{otherwise.} \end{aligned} \quad (5)$$

In this section, formula (5) will be extended to the situation where $J > 1$.

The function g in (2) is strictly convex for $x > 0$, and the constraints given by (4) are linear, so that the basic results in convex programming apply here without difficulty. That an optimal solution always exists was demonstrated by Kokan and Khan (1967). As above, denote the optimal solution by x^* . It follows from the Kuhn-Tucker Theorem (1951) that there exist $\lambda_j \geq 0$ such that

$$\nabla g(x^*) + \sum_{j=1}^J \lambda_j a_j = 0 \quad (6)$$

(∇ denotes the gradient) and

$$\lambda_j (a'_j x^* - 1) = 0 \quad (7)$$

for $j = 1, 2, \dots, J$. If $x > 0$ satisfies $\sum_{j=1}^J \lambda_j a'_j x \leq \sum_{j=1}^J \lambda_j$, then, combining (6) and (7),

$$-x' \nabla g(x^*) = \sum_{j=1}^J \lambda_j a'_j x \leq \sum_{j=1}^J \lambda_j = \sum_{j=1}^J \lambda_j a'_j x^* = -x^{*'} \nabla g(x^*). \quad (8)$$

By convexity, $g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*)$ (for all $x > 0$ with $x^* > 0$). Thus, from (8)

$$g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*) \geq 0.$$

It follows that x^* is the minimum of $g(x)$ subject to the conditions

$$\sum_{j=1}^J \lambda_j a'_j x \leq \sum_{j=1}^J \lambda_j \text{ for all } x > 0.$$

Since the minimization of g is unaffected by positive multiplicative constants, x^* also minimizes $g(x)$ subject to the constraints that $\sum_{j=1}^J \alpha_j^* a'_j x \leq 1$ and $x > 0$, where $\alpha_j^* = \lambda_j / \sum_{j=1}^J \lambda_j$.

The extension of formula (5) to an expression for the optimum multivariate allocation now consists of applying the former to the weighted sum $\sum_{j=1}^J \alpha_j^* a_j$:

$$x_i^* = \sqrt{c_i} / \left(\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}} \sum_{k=1}^I \sqrt{c_k \sum_{j=1}^J \alpha_j^* a_{kj}} \right) \quad \text{if } \sum_{j=1}^J \alpha_j^* a_{ij} > 0, 1 \leq i \leq I$$

$$= \infty \quad \text{otherwise.} \quad (9)$$

Notice that since x^* minimizes $g(x)$ subject to $a_j'x \leq 1$, with $x > 0$ for $1 \leq j \leq J$, it follows that mx^* minimizes $g(mx)$ subject to the constraints $a_j'(mx) \leq m$, with $x > 0$ for $1 \leq j \leq J$. Thus, as noted earlier, constraints on variances (or CV's) can be scaled by a factor m (or \sqrt{m}) if survey costs are too high.

Formula (9), of course, is computationally useful only if the α_j^* are known. However, this formula is useful for deriving the shadow prices and for developing an algorithm for obtaining x^* and the α_j^* .

4. SENSITIVITY OF SURVEY COST TO VARIANCE CONSTRAINTS

In many optimization problems, it is useful to know how the optimal solution behaves when the constraints are perturbed slightly. This can be especially true in survey research, where trade-offs between costs, survey operations and precision requirements are frequently required. In any case, the "shadow prices", given by $\partial g(x^*)/\partial v_k$, are useful in detecting small shifts in the variance constraints which could substantially reduce the overall survey cost.

Combining (2), (3), and (9), it is easily seen that the cost of the optimum allocation is

$$g(x^*) = \left(\sum_{i=1}^I \sqrt{c_i \sum_{j=1}^J \alpha_j^* a_{ij}} \right)^2 = \left(\sum_{i=1}^I \sqrt{c_i \sum_{j=1}^J \alpha_j^* W_i^2 S_{ij}^2 / v_j^2} \right)^2. \quad (10)$$

Thus

$$\begin{aligned} \frac{\partial g(x^*)}{\partial v_k} &= 2 \left(\sum_{i=1}^I \sqrt{c_i \sum_{j=1}^J \alpha_j^* W_i^2 S_{ij}^2 / v_j^2} \right) \sum_{i=1}^I \frac{-c_i \alpha_k^* W_i^2 S_{ik}^2 / v_k^3}{\sqrt{c_i \sum_{j=1}^J \alpha_j^* W_i^2 S_{ij}^2 / v_j^2}} \quad (11) \\ &= -2 \frac{\alpha_k^*}{v_k} \sqrt{g(x^*)} \sum_{i=1}^I \frac{a_{ik} \sqrt{c_i}}{\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}}} \\ &= -2 \frac{\alpha_k^*}{v_k} g(x^*) \sum_{i=1}^I a_{ik} \sqrt{c_i} / \left(\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}} \sum_{k=1}^I \sqrt{c_k \sum_{j=1}^J \alpha_j^* a_{kj}} \right) \\ &= -2 \frac{\alpha_k^*}{v_k} g(x^*) a_k' x^*. \end{aligned}$$

From (7) it follows necessarily that $\alpha_k^* = 0$ whenever $a'_k x^* < 1$, so that

$$\frac{\partial g(x^*)}{\partial v_k} = -2 \frac{\alpha_k^*}{v_k} g(x^*). \quad (12)$$

This formula is somewhat more complicated than the usual expression for shadow prices (e.g., see Luenberger 1984), due to the complex relationship between g and v_j .

Now consider increasing v_k by $(100\pi)\%$, $0 \leq \pi \leq 1$. Denote by $x^* + \Delta x^*$ the resulting perturbation in x^* . By (12),

$$g(x^* + \Delta x^*) - g(x^*) \approx \pi v_k \frac{\partial g(x^*)}{\partial v_k} = -2 \pi \alpha_k^* g(x^*). \quad (13)$$

Thus an increase of $(100\pi)\%$ in the k -th variance constraint results in a $(100)(2\pi\alpha_k^*)\%$ reduction in the overall survey cost.

5. PROGRAMMING CONSIDERATIONS

This section discusses some technical aspects of solving for x^* and gives a simple algorithm for finding both x^* and the coefficients α_j^* by searching over weighted averages $\sum_{j=1}^J \alpha_j a_j$. Define δ_{ij} by

$$\begin{aligned} \delta_{ij} &= 1 \text{ if } i = j \\ &= 0 \text{ if } i \neq j. \end{aligned}$$

For a vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_J)'$, define $\tilde{x}(\alpha)$ by

$$\begin{aligned} \tilde{x}_i(\alpha) &= \sqrt{c_i} / \left(\sqrt{\sum_{j=1}^J \alpha_j a_{ij}} \sum_{k=1}^I \sqrt{c_k \sum_{j=1}^J \alpha_j a_{kj}} \right) \text{ if } \sum_{j=1}^J \alpha_j a_{ij} > 0, 1 \leq i \leq I \\ &= \infty \text{ otherwise.} \end{aligned}$$

Notice that $\tilde{x}(\alpha^*) = x^*$. Now the iterative algorithm for finding x^* is defined as follows:

1. Take $\alpha_j^{(1)} = \delta_{1j}$, $1 \leq j \leq J$.
2. At step $n \geq 2$, find an index k for which

$$(a_k - a_j)' \tilde{x}(\alpha^{(n)}) \geq 0, 1 \leq j \leq J. \quad (14)$$

This gives the constraint which the current optimum solution violates by the largest margin. If $a'_k \tilde{x}(\alpha^{(n)}) \leq 1$, then terminate the algorithm. Otherwise, find $t^{(n)} \in (0, 1)$ for which

$$g(\tilde{x}(t^{(n)} \delta_k + (1 - t^{(n)}) \alpha^{(n)})) \geq g(\tilde{x}(t \delta_k + (1 - t) \alpha^{(n)})) \text{ for all } t \in [0, 1]. \quad (15)$$

3. Take $\alpha_j^{(n+1)} = t^{(n)} \delta_{kj} + (1 - t^{(n)}) \alpha_j^{(n)}$.
4. Terminate when $|\alpha_j^{(n+1)} - \alpha_j^{(n)}| < \epsilon$, $1 \leq j \leq J$, where ϵ is a predetermined convergence criterion.

To verify the convergence of the algorithm, first note that $\tilde{x}(\alpha)$ minimizes $g(x)$ subject to $\sum_{j=1}^J \alpha_j a_j' x \leq 1$. Thus, since $\sum_{j=1}^J \alpha_j a_j' x^* \leq \sum_{j=1}^J \alpha_j = 1$,

$$0 \leq g(\tilde{x}(\alpha^{(n)})) \leq g(x^*) \quad (16)$$

for all n . Furthermore, from (15), $g(\tilde{x}(\alpha^{(n)}))$ is nondecreasing, implying the convergence of $g(\tilde{x}(\alpha^{(n)}))$. To see that $\tilde{x}(\alpha^{(n)}) \rightarrow x^*$, first define

$$h_{k\alpha}(t) = \sum_{i=1}^I \sqrt{c_i \sum_{j=1}^J (t\delta_{kj} + (1-t)\alpha_j) a_{ij}} = \sqrt{g(\tilde{x}(t\delta_k + (1-t)\alpha))}. \quad (17)$$

Since $h_{k\alpha}(t)$ is concave (*i.e.*, $-h_{k\alpha}(t)$ is convex),

$$h_{k\alpha}(t) - h_{k\alpha}(0) = t h'(0) + O(t^2) \quad (18)$$

$$\begin{aligned} &= -t \sum_{i=1}^I \frac{\sum_{j=1}^J (\delta_{kj} - \alpha_j) a_{kj} \sqrt{c_k}}{2\sqrt{c_i \sum_{j=1}^J \alpha_j a_{ij}}} + O(t^2) \\ &= (t/2) \sqrt{g(\tilde{x}(\alpha))} (a_k' \tilde{x}(\alpha) - 1) + O(t^2). \end{aligned}$$

By allowing t to tend toward zero, it follows that there exists $t \in (0,1)$ for which

$$\sqrt{g(\tilde{x}(t\delta_k + (1-t)\alpha))} = h_{k\alpha}(t) > h_{k\alpha}(0) = \sqrt{g(\tilde{x}(\alpha))}$$

if and only if $a_k' \tilde{x}(\alpha) > 1$. Thus it follows from (15) that the constraints are satisfied at convergence; combining this with (16) implies that $\lim_{n \rightarrow \infty} \tilde{x}(\alpha^{(n)}) = x^*$.

In carrying out the algorithm, Step 2 requires a search for $t^{(n)}$. Define $h_{k\alpha}(t)$ as in (17). It is clear from the preceding discussion that $a_k' \tilde{x}(t\delta_k + (1-t)\alpha^{(n)}) = 1$ when $h(t)$ (and hence g) is at a maximum. Furthermore, since $h_{k\alpha}(t)$ is strictly concave, $h'_{k\alpha}(t)$ is nonincreasing in t and thus the point where $h'_{k\alpha}(t) = 0$ is unique. It follows that a binary search for the point where $h_{k\alpha}(t)$ is maximized can be implemented by simply checking to see whether $a_k' \tilde{x}(t\delta_k + (1-t)\alpha^{(n)}) = 1$, providing a rapid means of obtaining a close approximation for $t^{(n)}$.

As described above, the algorithm takes a_1 as the initial value. This is completely arbitrary, since any of the a_j , $1 \leq j \leq J$, would do. In practice, the constraint for which the optimum allocation (*i.e.*, formula (5)) yields the highest cost is generally a good choice for the starting value.

Notice that Step 2 of the algorithm will require IJ calculations in formula (14) and a 10-step (say) binary search of $3I + J + 1$ calculations each in formula (15), while J calculations must be carried out in Step 3. Thus each iteration of the algorithm is $O(IJ)$. From (18), at the n -th iteration,

$$h'_{k\alpha}(0) \approx \frac{1}{2} h_{k\alpha}(0) (a'_k \tilde{x}(\alpha^{(n)}) - 1)$$

so that $a'_k \tilde{x}(\alpha^{(n)})$ is approximately proportionate to $h'_{k\alpha}(0)$ (up to an additive constant). Heuristically, $h'_{k\alpha}(0)$ is the “slope” of h in the direction of a_k , suggesting that the algorithm is essentially a steepest descent (or ascent, in this case) procedure. This, in turn, suggests a linear rate of convergence (see, for example, Forsyth 1968).

In the author’s experience (see Bethel 1985), the algorithm converges quickly for most moderately sized problems. For example, sample allocation problems with 20-30 strata and 5-10 constraints were solved in 3-5 seconds using the algorithm (on a Compaq 38620 with a 30387 math co-processor) versus 6-8 seconds using a sequential unconstrained minimization technique (SUMT) implementing a penalized steepest descent algorithm. Run times vary considerably depending on the magnitude of the problem, the number of active constraints, and, obviously, machine characteristics. The author’s computing experience (with problems of 20-30 strata and 5-10 constraints) includes the Macintosh SE (30 seconds to 2 or 3 minutes), Leading Edge Model D (1 to 5 minutes), Zilog System 8000 (5 to 60 seconds), and the Compaq mentioned above (5 to 10 seconds). However, the run times are generally insignificant in comparison with the labor involved in creating files and other preparatory tasks. In particular, it may take several hours to find an acceptable starting value for the SUMT algorithm. Thus a strong feature of the algorithm described in Steps 1-4 above is that it requires no external initial values. Moreover, it is relatively easy to program, requiring only 40 or 50 lines of code.

An even simpler algorithm is given by Chromy (1987). It can be adapted to our notation and general approach as follows: Set $\alpha_j^{(1)} \equiv 1/J$, and, for $n \geq 2$, let

$$\alpha_j^{(n)} = \alpha_j^{(n-1)} (a'_j \tilde{x}(\alpha^{(n-1)}))^2 / \sum_{j=1}^J \alpha_j^{(n-1)} (a'_j \tilde{x}(\alpha^{(n-1)}))^2 \quad 1 \leq j \leq J. \tag{19}$$

Like the algorithm described in steps 1-4 above, (19) requires no external initial values; (19), however, requires even less programming effort and, based on several comparisons, it appears to converge considerably more quickly. Unfortunately, there is apparently no formal proof of convergence, although considerable practical experience (see Chromy 1987 for a more detailed discussion) suggests that it has good convergence properties.

6. EXAMPLE

Tables 1-3 present an example drawn from a survey of commercial establishments. (Only the strata for educational institutions are shown here.) Four of the primary variables of interest are given: area of enclosed floorspace, age of building, number of full-time employees, and percent of buildings heated by oil. Table 1 gives the stratum level variance information. Here the standardized precision units are computed as

$$a_{ij} = \frac{W_i^2 S_{ij}^2}{\bar{Y}_j^2 v_j^2}$$

Table 1.
Allocation Example: Survey of Educational Institutions.

Stratum Standard Deviation					
	Weight	Floorspace	Age	Employees	Pct. Oil Heating
Stratum					
1	.5158	22,319.11	43.71	25.72	48.15
2	.2632	24,056.21	16.68	27.09	36.79
3	.1184	54,201.75	24.70	17.11	48.04
4	.0711	155,514.21	16.01	59.46	38.07
5	.0184	125,239.21	14.74	51.27	48.80
6	.0132	355,392.69	20.90	212.13	57.74
Mean:		54,641.85	43.03	45.23	67.58
v_k :		.06	.06	.06	.06
Standardized Precision Units					
		Floorspace	Age	Employees	Pct. Oil Heating
Stratum					
1		12.33	76.24	23.90	37.52
2		3.73	2.89	6.93	5.70
3		3.83	1.28	.56	1.96
4		11.36	.19	2.44	.45
5		7.37	.01	.12	.05
6		2.03	.01	1.06	.04
Required					
Sample Size:		222	149	127	121

where $v_j = .06$ for all variables (so that the half-width of a 90% confidence interval will be approximately 10% of the mean). Also given are the sample sizes required for Neyman allocation for each of the variables taken individually. Survey costs are assumed to be constant across strata.

Table 2 gives the first-pass solution, which requires a sample of 241 units. The normalized Lagrangian coefficients and the achieved precision levels are given, from which it is apparent that floorspace and building age are dominating the solution while the other variables are not “active”. Here the starting value $\alpha^{(1)} = (1,0,0,0)$ was used; because the third and fourth constraints were always satisfied, there was only one iteration with a 9-step binary search for $t^{(1)}$. (The successive estimates for the optimal t were 1/2, 1/4, 3/8, 5/16, 11/32, 21/64, 43/128, 85/256, and 171/512.) Also given in Table 2 are the 10% shadow prices: 10% increases in the first (or second) constraints would result in a sample size reduction of approximately 32 (or 16) units. Since the third and fourth constraints are not active in the solution, changing their CV requirements would have no effect on the allocation or the sampling costs.

Table 3 gives a second pass solution under the requirement that the total sample size is no larger than 200. The optimal solutions are thus scaled by 241/200 (so that the optimal allocation goes down by 200/241) and the resulting CV’s are scaled by $\sqrt{241/200}$. The new 10% shadow prices are -27 and -13 for the first and second constraints, reflecting the decrease in the overall survey cost. Notice that there is approximately a 10% increase in the CV’s (from the original ones in Table 1), so that the sample reduction of 48 predicted by the shadow prices in Table 2 compares favorably with the actual 41 unit reduction. (The shadow price predictions will always be somewhat optimistic due to the linear approximation.)

Table 2.
Allocation Example: First Pass Optimum Solution.

Stratum	$\sum \alpha_j^* a_{ji}$	x_i^*	Optimum Allocation	
1	33.6749	.0111	90	
2	3.4495	.0347	29	
3	2.9783	.0373	27	
4	7.6294	.0233	43	
5	4.9119	.0291	34	
6	1.3554	.0553	18	
Total:			241	
	Floorspace	Age	Employees	Pct. Oil Heating
Lagrangian Multiplier (Normalized):	.6660	.3340	.0000	.0000
Achieved Precision:	.0600	.0600	.0481	.0502
10% Shadow Prices:	-32	-16	0	0

Table 3.
Allocation Example: Optimum Solution for Sample Size Limited to 200.

Stratum	$\sum \alpha_j^* a_{ji}$	x_i^*	Optimum Allocation	
1	33.6749	.0134	75	
2	3.4495	.0418	24	
3	2.9783	.0449	22	
4	7.6294	.0281	36	
5	4.9119	.0351	29	
6	1.3554	.0666	15	
Total: 201				
	Floorspace	Age	Employees	Pct. Oil Heating
Lagrangian Multiplier (Normalized):	.6660	.3340	.0000	.0000
Achieved Precision:	.0657	.0658	.0528	.0551
10% Shadow Prices:	-27	-13	0	0

7. DISCUSSION

In this paper we have given a formal representation for the optimal sample allocation for a multipurpose survey with linear variance constraints, and derived expressions for the partial derivatives of the cost function with respect to the precision constraints. The latter result, in particular, provides approximations that are useful in survey planning, permitting a great deal of exploratory work without exact computer calculations.

Throughout the paper, the normalized Lagrangian multipliers, α_j^* , play a key role. In particular, we have noted that whenever the j -th variance constraint is not "active" in the solution to the allocation problem, the j -th Lagrangian $\alpha_j^* = 0$.

The optimization approach discussed in this article yields a continuous solution, which must then be rounded in some way to provide integer stratum sample sizes. Clearly this rounding will cause some deviation from optimality. However, the objective function here is generally considered to be rather insensitive to small deviations from optimality (see Cochran 1977), so that exact integer solutions are probably not cost effective. In fact, it seems likely that round-off error would be insignificant in comparison with the sampling errors in estimates of means and variances that would normally be available for developing an optimized survey design.

Finally the reader will recall that finite population correction factors have been ignored throughout this paper. It is easy to include these in the allocation model by manipulating equations (1) and (3), although that would cause equation (13) to be somewhat imprecise. However, it should be kept in mind that even when the FPC is non-negligible for some of the strata, the overall effect usually is negligible. In any case, the FPC term, $\sum_{i=1}^I W_i^2 S_{ij}^2 / N_i$, can always be calculated in order to evaluate the situation and, if necessary, it can be added to v_k in formula (13) to obtain exact results.

ACKNOWLEDGEMENTS

The author acknowledges the comments of the referees and useful discussions with Rick Williams, of Research Triangle Institute, and Patrick McCarthy, of Applied Management Sciences, all of which led to substantial improvements in the article.

REFERENCES

- ARTHANARI, T.S., and DODGE, Y., (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.
- BETHEL, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Section, American Statistical Association*, 209-212.
- CHATTERJEE, S. (1968). Multivariate stratified surveys. *Journal of the American Statistical Association*, 63, 530-534.
- CHATTERJEE, S. (1972). A study of optimum allocation in multivariate stratified surveys. *Skandinavisk Actuarietidskrift*, 55, 73-80.
- CHROMY, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Section, American Statistical Association*, 194-199.
- DALENIUS, T. (1953). The multivariate sampling problem. *Skandinavisk Actuarietidskrift*, 36, 92-102.
- DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wicksall.

- FOLKS, J.L., and ANTLE, C.E. (1965). Optimum allocation of sampling units when there are R responses of interest. *Journal of the American Statistical Association*, 60, 225-233.
- FORSYTH, G.E. (1968). On the asymptotic directions of the s -dimensional optimum gradient method. *Numerische Mathematik*, 11, 57-76.
- HARTLEY, H.O. (1965). Multiple purpose optimum allocation in stratified sampling. *Proceedings of the Social Statistics Section, American Statistical Association*, 258-261.
- HUDDLESTON, H.F., CLAYPOOL, P.L., and HOCKING, R.R. (1970). Optimum sample allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society A.*, 139, 80-95.
- KOKAN, A.R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society A.*, 126, 557-565.
- KOKAN, A.R., and KHAN, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society B.*, 29, 115-125.
- KUHN, H.W., and TUCKER, A.W. (1951). Nonlinear programming. *Proceedings 2nd Berkeley Symposium Mathematical Statistics and Probability*.
- LUENBERGER, D.G. (1984). *Linear and Nonlinear Programming*. Reading, Massachusetts: Addison-Wesley.
- NEYMAN, J. (1934). On the two different aspects of the representative method: the method of representative sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, 558-625.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin and Company.

The Role of Demographic Factors in the Analysis of Survey Versus Diary Purchase Reporting Accuracy

EDWARD R. BRUNING and MICHAEL Y. HU¹

ABSTRACT

In this article the authors evaluate the relative performance of survey and diary data collection methods in the context of the long-distance telephone communication market. Based on an analysis of 1,530 respondents, the results indicate that two demographic variables, sex and income, are important in explaining the difference in survey reporting and diary recording of usage data.

KEY WORDS: Survey; Diary; Data collection.

1. INTRODUCTION

A perusal of the marketing literature underscores our lack of knowledge regarding the relative accuracy of survey and diary methods for collecting consumer expenditure data. Clearly, the resolution of this issue has ramifications for researchers as well as those for whom the research is conducted. Wind and Lerner (1979) stress the need to appropriately evaluate the two methods and to identify the characteristics of those reporting purchase behavior accurately versus those that have a high discrepancy between reported and actual consumption. To be sure, an analysis of the discrepancy focuses attention on the data collection instrument, for the choice of instrument could affect management decisions relating to "product positioning and market segmentation strategies, advertising media and copy research, and concept/product testing." (Wind and Lerner 1979).

The purpose of our article is to assess empirically the relationship between several demographic variables and the two expenditure reporting methods from a single sample of respondents in the U.S. long-distance telephone market. We present additional evidence on the issue initially posed by Wind and Lerner. First, the current state of knowledge regarding the nature of the two instruments is surveyed. Then the research methodology is described and findings from the long-distance telephone market are reported. We conclude with a number of implications relevant to both providers and users of consumer expenditure data.

2. LITERATURE REVIEW

The two prominent methods for recording household consumption expenditures are survey (recall) methods, whereby household members are asked to recall expenditures made during a predefined period, and the diary method, whereby a daily or weekly log is maintained which identifies specific expenditures. Neter (1970) provides case examples and empirical studies which address the relative advantages and disadvantages of the two expenditure collection devices but do not compare their relative accuracies. In general, the survey approach possesses advantages in economy while simultaneously possessing a number of disadvantages relative to the diary method. Because of time and resource constraints, most researchers utilize the survey method even with the multitude of measurement problems.

¹ Edward R. Bruning and Michael Y. Hu, Graduate School of Management, Kent State University, Kent, Ohio, 44242.

It is commonly believed that diary methods have advantages over survey approaches principally because diarists have the opportunity to record the event within a short period after it has occurred. For this reason, Sudman and Ferber (1971) have all but discredited the survey approach for collecting expenditure data and have suggested the exclusive use of diaries. But the diary method is not problem free. The authors evaluated households in the Chicago area in 1972 and found evidence of underreporting by the survey method with respect to the number of purchases. They also found that respondents had difficulties in separating purchases into specific item categories with the survey recall method.

A number of writers report that the diary approach is appropriate only for certain expenditure categories (Pearl 1968; Grooteart 1986; Wind and Lerner 1979; Stanton and Tucci 1982). Pearl (1968) has stated that individual diaries are to be preferred because of reporting thoroughness. For large ticket items the method is preferred; however, reporting frequency declines for small valued purchases. Grooteart (1986) adds to this prescription by suggesting that all eligible household members keep diaries to reduce omissions in expenditure reporting. Wind and Lerner (1979) and Stanton and Tucci (1982), in separate studies on expenditure reporting for specific food items, substantiate the superiority of the panel method relative to surveys.

The construction and design of the diary instrument poses collection problems (Kemsley 1961; Kemsley and Nicholson 1960; Lewis 1948; Sudman 1964a, b; Sudman and Ferber 1971; Walsh 1977). Kemsley (1961) and Kemsley and Nicholson (1960) evaluated record books kept on consumer expenditures over a three week period in 1953. They found that significant variations occurred in expenditure recording over the three week period by type of expenditure and by season of the year. Lewis (1948) evaluated the accuracy of weekly versus monthly diary recording of grocery and clothing expenditures. The author found a 16% reduction in monthly reporting in comparison to weekly expenditure reporting. Sudman (1964a) and Sudman and Ferber (1974) studied alternative means of obtaining consumer expenditure data. They evaluated the role of compensation, training of respondents, and method of reporting. In the studies they conducted, compensation was significant in improving respondent cooperation and accuracy, and direct training aided in respondent reporting accuracy. The frequency of purchase and the construction of the reporting form were also important in reporting accuracy.

Other studies have focused more explicitly on consuming unit cooperation. (Kemsley and Nicholson 1960; Pearl 1968; Sudman and Ferber 1974). Kemsley and Nicholson (1960) report that the size of the individual purchase has a significant effect upon the degree to which respondents cooperate in reporting expenditures. Pearl (1968) and Sudman and Ferber (1974) emphasize the incentive payments in terms of amount and duration in generating cooperative expenditure reporting.

An additional concern with the diary method is the extent of panel mortality (Sandage 1956; Sodol 1959; Sudman 1964a, b) and panel decay (McKenzie 1983; Sandage 1956; Sodol 1959; Sudman 1964 a, b). Sandage (1956) investigated whether consumer panels develop bias as a result of being interviewed. Based on three separate investigations on Indiana farm households over the period 1947-1954, the author found that bias was not a significant concern with panel collection methods. Sudman (1964a, b), however, found mortality tended to be greater for male respondents. In addition, the degree of effort involved in recording appeared to have no impact on accuracy or mortality rates for respondents involved in panel recording of expenditures. In terms of panel decay, McKenzie (1983) reported that greater attrition occurred with longer panel periods while Sandage (1956) found that repeated use of a given panel did not result in a bias in reporting accuracy.

Parfitt (1967) argues that housewives in surveys recall accurately only purchases for frequently bought products in the most recent past. Thus, the diary recording of past purchases

yields a more reliable and accurate measure than survey reporting. In surveys, respondents typically are asked to report purchases over a long time period or to engage in a mental averaging exercise to arrive at an expenditure figure for a typical week or month. As a consequence, Parfitt (1967) concludes that a strong likelihood exists for respondents to exaggerate the amount and frequency of purchases and to oversimplify the complexity of the expenditure decision.

As indicated in an earlier section of this paper, our research focuses on the accuracy of survey versus diary purchase reporting. Only a few articles address this issue empirically. Wind and Lerner (1979) analyze the validity of survey versus diary approaches in accounting for consumer expenditures. Their data are taken from a sample of 450 housewives serving on a MRCA consumer diary panel. The housewives completed a mail survey questionnaire and were instructed to maintain a record of their expenditures of various brands of margarine for a six month period. The results indicate a discrepancy in the relative accuracy of the two reporting methods between the aggregate and the individual consumer response level. At the aggregate level, survey and diary instruments are consistent in predicting the rank-ordering of brand market shares. Major discrepancies are detected, however, at the consumer level as survey responses are less accurate as compared to diary reporting. The authors attribute this inaccuracy as resulting from ignorance, forgetfulness, poor survey questioning, reporting errors, falsification, and interviewer bias.

Stanton and Tucci (1982), following the work reported by Wind and Lerner (1979), sample 7,945 participants in the National Food Consumption Survey (1977-78). Personal interviews are used as the reporting vehicle for food expenditures which occurred in the previous twenty-four hour period. The participants were asked to maintain diaries of all food and beverage expenditures for two days following the interview. Their results indicate that, at aggregate levels, personal interviews provide information which is as accurate and reliable as diary reports. They were not able to address the relative accuracy of the two approaches at the consumer level because of the nature of the data.

The apparent discrepancies in results reported by Wind and Lerner (1979) and Stanton and Tucci (1982) may be attributable to the differences in the time frames within which consumers operated in reporting expenditures. In Wind and Lerner's study, respondents were requested to report the brand most often purchased. Questions of this nature require a greater amount of recall since the time reference is over an extended period. In Stanton and Tucci's study, however, the reporting period is restricted to the previous twenty-four hours. Parfitt (1967) indicates that respondents are more effective in reporting recent purchases. In this light, Stanton and Tucci's conclusion is not truly surprising and, furthermore, does not contradict the results of Wind and Lerner's analysis since recall for both the survey and diary recording methods was high.

3. THE STUDY

During the years 1978 and 1979, AT&T (American Telephone and Telegraph Company) initiated a major data collection effort with the objective of providing information for corporate market planning and strategy formulation in its residential long-distance telephone market. A nationally projectable sample of roughly 4,000 households were recruited and asked to participate on a panel for a period of twelve months. The sample was demographically balanced with respect to six variables: population density, income, marital status, age, sex and geographical region of domicile.

The entire panel responded to a pre-assessment survey instrument administered through the mail in January 1978. Once completed, each panel member was instructed to fill out a weekly diary over the next twelve months. In the pre-assessment phase, respondents were asked to

respond to the question: "During an average or typical month, how often do you communicate for non-business reasons with relatives and friends who reside at least 50 miles from your home?" This measure is referred to as [PERCEIVED 1]. Also, each panel member in the pre-assessment phase responded to the question: "Would you consider yourself a heavy, medium, light or non-user of long distance calling?" (Refer to this measure as PERCEIVED 2). So that comparisons could be made between panel and survey data collection methods, panel respondents were asked to record information on the frequencies of long-distance communication by day of the week. This measure is referred to as REPORTED 1.

Throughout the entire study every attempt was made to conceal the sponsor of the project. Moreover, the positions of the response categories were randomized in order to remove any possibility of position bias. A sample of 2,350 respondents was retained after twelve months of reporting. Panel attrition was perceived to be a potential problem in this study because attrition rates may vary substantially among demographically defined subgroups. In order to resolve this problem, a sample balance program was developed and used to randomly select a subsample of participants from the pool of 2,350 respondents which would be demographically balanced. After editing and sample balancing, 1,530 panel members who had completed the pre-assessment and the twelve-month diaries were used in this study.

4. DATA ANALYSIS

An important question in the pre-assessment survey asked the respondents to report their "perceived" usage for a typical month [PERCEIVED 1]. In order to obtain consistency in the unit of measurement, weekly diary recorded usage [REPORTED 1] is aggregated to twelve monthly totals for each respondent. Refer to the aggregated diary reported measure as REPORTED 2. Matched differences between "perceived" usage reported in the pre-assessment survey [PERCEIVED 1] and "actual" usage extracted from diaries [REPORTED 2] are calculated for each respondent for twelve monthly periods as well as for the average of the twelve months. A one-way ANOVA design is employed monthly and for the twelve month average to detect if significant variations exist with respect to the matched differences across levels of several demographic variables: sex, income, education, and age. An a posteriori contrast test is performed to compare all possible pairs of level means for each demographic variable. Finally, to evaluate the effects of interactions among the four demographic variables, a four-way ANOVA procedure is employed using the twelve-month average scores.

5. RESULTS

5.1 Survey and Diary Average Reported Usage

Table 1 reports the average number of long distance telephone communications extracted from respondent diaries for each of the twelve months as well as the usage for a typical month [PERCEIVED 1] taken from the pre-assessment survey. Interestingly, this "perceived" usage reported in the pre-assessment survey is substantially greater than actual recorded usage [REPORTED 2] for each month of the analysis.

The diary averages indicate the presence of seasonality in the usage. December 1978 usage of 4.123 is the highest among the twelve reported months. Even though the pre-assessment survey requested the respondents to report usage for an average or typical month, it is quite likely that they would use December 1977 as the basis for response since the pre-assessment

survey was administered in January 1978. A one-sample t-test indicates that the average of the paired-difference between pre-assessment and the December diary usage, 0.235, is significantly different from zero (p -value = 0.001). By the same token, t-test results for the other eleven averages are statistically significant. These results imply that the respondents have indeed over-estimated in the pre-assessment survey as compared to the diary reported usage.

A potential concern is that the reported usage in the pre-assessment survey could be influenced by the unusually high usage in December 1977. If so, then it is argued that the results of our study are subject to seasonality bias. In addressing this issue, the authors have examined the difference between the reported usage in the pre-assessment survey and the December 1978 diary. Comparing the same months over a year of time could help to eliminate the seasonality factor. As indicated in Table 1, this difference is statistically significant. This difference, however, can be due to the difference in the data collection method and to a trend factor since the comparison involves two different years. Assuming a positive trend in the usage of services over time, the reported usage in December 1978 should be higher than that of December 1977. The data from Table 1 indicates quite the contrary. Usage in December 1977 was significantly higher than that reflected in December 1978. Thus, this evidence leads us to conclude that there is indeed a significant difference due to the data collection method. Respondents in our study had over-estimated their usage in the pre-assessment survey as compared to their estimates reported in the diary.

Prior to our analyzing the relationship between the difference in survey versus diary data collection methods and the several demographic variables, it is important to evaluate the role played by actual usage in explaining this difference. Our reasoning for this test is that if the difference between survey reporting and diary recording is due to the absolute level of usage, then further analysis would prove suspect since experience (learning) would tend to bias our dependent variable (McKenzie 1983). On the other hand, if no statistical significance is attributable to the differences in collection methods and absolute usage levels, then the analysis with the demographic variables would be of greater validity.

Table 1
Average Absolute Number of Long-Distance Telephone
Communications and Pre-assessment Survey Estimates

Month	Average Absolute Number of Communications
February	3.516
March	3.878
April	3.486
May	3.610
June	3.414
July	3.604
August	3.606
September	3.250
October	3.426
November	3.518
December	4.123
January	3.891
Preassessment Survey Estimate	4.358
	$n = 1530$

Table 2
One-Way ANOVA Results Relating the Degree of Long-Distance Telephone Usage and the Difference Between Survey and Diary Reporting (12 month average)

	Degree of Usage				P-Value
	Heavy	Medium	Light	Non-User	
Mean Difference (survey-diary)	0.762	0.799	0.795	0.580	0.9905
<i>n</i>	316	605	547	45	

Table 2 reports the results of the analysis of the relationship between the difference in survey [PERCEIVED 1] and diary [REPORTED 2] reportings and the degree of absolute usage [PERCEIVED 2]. McKenzie evaluated the form of both response and recording bias involving the collection of telephone call details by diary methods. Response rate was found to vary with customer usage. Furthermore, telephone usage recording rates tended to decrease with usage as well. Thus, telephone call data collected by diary methods are subject to several biases. Our study focuses on the difference in survey versus diary collected data and customer usage where the emphasis lies with the discrepancy between “perceived” and “actual” consumption/purchase and the level (degree) of usage. Even though recording biases exist with both methods, nonetheless, the *difference* between the two recordings is not related to usage.

In addition, the validity of using PERCEIVED 2 as a categorization variable can be examined by correlating this measure with REPORTED 2 and PERCEIVED 1. REPORTED 2 and PERCEIVED 1 measurements were first categorized into heavy, medium, light, and non-user employing different cut-off levels. Cross-tabulations were then conducted between PERCEIVED 1 and these two categorical measures. Significant statistical relationships were detected in all cases.

Our dependent variable is the difference in the survey [PERCEIVED 1] and diary usage recordings [REPORTED 2] and the independent variable is the degree of usage divided into four levels: heavy, medium, light and non-user [PERCEIVED 2]. The results from the one-way ANOVA procedure using the least-squares estimation procedure indicate that the degree of usage is not statistically significant ($p = .9905$) in explaining the recorded usage difference between the survey and diary methods. A one sample t-test of each of the four individual group means showed that each mean was statistically different from zero at the 0.01 significance level. Therefore, the results imply that with respect to each of the four usage groups the positive mean values represent that respondents tend to over-estimate usage in the pre-assessment survey relative to the diary recording method.

5.2 Relationship Between Survey and Diary Reported Usage Differences and Selected Demographic Variables

In Table 1 we reported the existence of a substantial difference between survey and diary collection methods for the same respondents over a twelve month period. An interesting question is: what accounts for the perceptual bias in survey reporting of purchase data? To answer this question, a number of demographic factors are evaluated. Several levels of each factor are specified and a one-way ANOVA procedure is employed to account for the reporting differences. Tables 3 through 7 report the results of the analyses.

Table 3
One-Way ANOVA Results Relating Sex of Respondent and
the Difference Between Survey and Diary Reporting of Data

Month	Differences by Sex (Survey — Diary)		ANOVA p-value
	Male	Female	
February	0.412	1.135	0.006*
March	−0.015	0.818	0.005*
April	0.379	1.201	0.002*
May	0.310	1.304	0.008*
June	0.562	1.205	0.016**
July	0.376	1.008	0.018**
August	0.395	0.987	0.031**
September	0.927	1.225	0.258
October	0.605	1.149	0.042**
November	0.593	1.003	0.129
December	−0.112	0.464	0.041**
January	0.164	0.675	0.075**
Mean ^a	0.380	0.990	0.010*
n	617	911	

^a Twelve month average.
* Significant at the 0.01 level.
** Significant at the 0.05 level.

5.3 Sex

The relationship between the difference in survey and diary recordings of usage and sex of the respondent is depicted in Table 3. The one-way ANOVA p-values are statistically significant for 9 of the 12 months at the 0.05 level or below and significant at the 0.01 level for the twelve month average. Thus the results indicate that both male and female respondents over-estimate their actual usage of long distance telephone service and that females over-estimate to a greater degree than do males.

5.4 Income

In Table 4 we present the difference between survey and diary usage reports in relation to respondents' household income level. For 6 of the 12 months the one-way ANOVA p-values are statistically significant at the 0.05 level or better and the 12-month average is significant at the 0.037 level. Furthermore, the results of Tukey's Studentized t-test indicate that respondents with annual household income in the Category 1 range (\$5,000 or less) are statistically distinct from respondents earning incomes within the range of \$10,001 to \$20,000.

An obvious anomaly in the findings reported in Table 4 is that for respondents within the lowest income category (\$5,000 or less), estimated average monthly usage is below the actual monthly usage in 9 of the 12 periods. Furthermore, with increasing household income a definite persistence to over-estimate usage occurs although this process begins to subside at the highest income category. At lower income levels consumers may perceive long-distance telephone service as a luxury item with respect to the other modes as well as with regard to other consumer expenditures. Consequently, when asked to report expected usage, as in a survey, respondents from this income strata tend to discount their perceived usage because of the belief

Table 4
ANOVA Results Relating Respondent's Income Level and
the Difference Between Survey and
Diary Reporting of Long-Distance Telephone Usage

	Differences by Income (Survey — Diary)					<i>p</i> -value
	0-\$5,000 (1)	\$5,001 – 10,000 (2)	\$10,001 – 15,000 (3)	\$15,001 – 20,000 (4)	Over 20,000 (5)	
February	–0.010	–0.583	1.180	1.120	0.571	0.110
March	–0.480	–0.738	0.780	1.009	–0.062	0.019**
April	–0.337	0.851	1.188	1.258	0.550	0.031
May	–0.327	0.560	0.928	0.991	0.636	0.220
June	0.102	0.911	1.027	1.331	0.756	0.249
July	–0.439	0.500	0.895	1.050	0.694	0.128
August	–0.408	0.512	1.021	1.235	0.498	0.036**
September	0.306	0.798	1.298	1.367	0.976	0.301
October	–0.469	0.542	1.231	1.413	0.720	0.009*
November	0.010	0.494	0.941	1.214	0.741	0.248
December	–1.010	0.060	0.209	0.792	0.101	0.050**
January	–0.633	–0.339	0.654	0.956	0.392	0.030**
Mean ^{a,b}	–0.308	0.517	0.946	1.145	0.548	0.037**
<i>n</i>	98	168	373	341	536	

^a Twelve month average.

^b Tukey's Contrast Test: (1) and (4) and (1) and (3) are different at the $p = 0.05$ level.

* Significant at the 0.01 level.

** Significant at the 0.05 level.

that limited monies should be spent elsewhere. At the actual point of consumption, however, relative values may have changed since the urgency of the situation may dictate a long-distance telephone call is indeed the low-cost option relative to alternative communication means. Thus, survey reporting of planned usage may deviate from diary recordings of actual usage because of situational factors that intervene during the time of consumption.

As respondents' household incomes increase, long-distance telephone use is still perceived as a superior good; however, whereas respondents in lower income levels perceive long-distance telephone use as an expendable (and perhaps frivolous) purchase, wealthier respondents "expect" to employ the telephone more often than the other modes. Thus, when surveyed as to their "expected" usage, wealthier respondents tend to overestimate the number of long-distance telephone communications since in most situations it is their preferred method of communicating.

5.5 Age

Table 5 reports that respondents at every age level tend to over-estimate their "perceived" usage relative to "actual" usage as recorded in diaries. Although the one-way ANOVA p -values indicate the nonexistence of a significant relationship between measurement methods and age of the respondent; nonetheless, in 10 of the 12 months respondents less than 31 years of age incurred the lowest difference relative to older respondents. In addition, average differences for respondents between 31 and 40 and over 50 were lower than the average for the less than

Table 5
One-Way ANOVA Results Relating Respondent's Age and
the Difference Between Survey Reporting and
Diary Recording of Long-Distance Telephone Usage

Month	Differences by Age (Survey — Diary)				<i>p</i> -value
	Below 31 (1)	31-40 (2)	41-50 (3)	Over 50 (4)	
February	0.632	0.749	1.026	0.949	0.310
March	0.016	0.348	0.837	0.709	0.210
April	0.413	1.083	1.174	0.889	0.209
May	0.305	0.706	1.085	0.923	0.217
June	0.525	0.845	1.570	0.989	0.080
July	0.535	0.706	1.226	0.667	0.371
August	0.507	0.807	1.070	0.667	0.578
September	0.924	1.003	1.459	1.109	0.580
October	0.789	0.816	1.307	0.903	0.583
November	0.632	0.805	1.415	0.741	0.240
December	0.337	0.203	0.574	-0.030	0.494
January	0.603	0.519	0.922	0.069	0.197
Mean ^a	0.518	0.716	1.139	0.715	0.385
<i>n</i>	383	374	270	495	

^a Twelve month average.
* Significant at the 0.01 level.
** Significant at the 0.05 level.

31 group in each of the twelve periods. Thus, the relationship between differences in survey and diary usage reports and age of respondent is a monotonically increasing function up to age 50 where the difference, although still positive, declines after age 50. Again, the average differences across the various age levels are not statistically significant based on the one-way ANOVA or the Tukey Studentized *t*-tests.

5.6 Education

The relationship between the difference in survey versus diary reported usage and respondents' level of education is depicted in the statistics found in Table 6. As reported in the table, a general tendency to over-estimate usage in surveys is characteristic of respondents at all education levels. Respondents with the least amount of formal education tend to over-estimate usage in survey reporting to a lesser extent than respondents with more formal education. The greatest tendency to over-estimate usage in surveys occurs for respondents who have completed high school followed by those who have had some college. The results of the one-way ANOVA and Tukey Studentized *t*-tests, however, indicate that the differences across education levels are not statistically significant at the *p* = 0.05 level.

5.7 Four-Way ANOVA Results

The main and interaction effects of the demographic variables as explanations of the difference between survey and diary purchase data reporting are presented in Table 7. It is reported that income, sex, and their interaction are the variables with statistically significant *p*-values. All other main and interaction affects are insignificant in explaining variations in the difference variable.

Table 6
One-Way Anova Results Relating Respondent's Education Level and the
Difference Between Survey and Diary Reported Usage

Month	Differences by Education (Survey-Diary)				<i>p</i> -Value
	Some High School (1)	Completed High School (2)	Some College (3)	Completed 4-Yr. Col. Deg (4)	
February	0.790	1.015	0.951	0.578	0.546
March	0.290	0.853	0.592	0.059	0.117
April	0.556	1.275	0.979	0.445	0.078
May	0.685	1.134	0.756	0.345	0.139
June	0.548	1.158	1.111	0.696	0.368
July	0.194	0.931	1.021	0.467	0.195
August	0.347	0.891	0.845	0.620	0.681
September	1.040	1.137	1.190	1.018	0.959
October	0.468	1.195	0.826	0.878	0.475
November	0.508	1.119	0.896	0.592	0.383
December	0.081	0.500	0.244	0.061	0.558
January	-0.097	0.626	0.842	0.129	0.138
Mean ^a	0.438	0.986	0.854	0.491	0.294
<i>n</i>	124	476	431	490	

^a Twelve month average.

* Significant at the 0.01 level.

** Significant at the 0.05 level.

Table 7
Four-Way ANOVA Results Relating
Demographics (Sex, Education, Age, and Income)
to the Differences Between Survey and
Diary Recording of Long-Distance Telephone Usage

Variable	Df	Sum of Squares	F-Value	<i>p</i> -Value
Sex	1	131.082	6.48	0.011**
Education	3	79.001	1.30	0.272
Age	3	58.465	0.96	0.409
Income	4	210.077	2.60	0.035**
Sex & Education	3	77.629	1.28	0.280
Sex & Education	4	220.032	2.72	0.028**
Sex & Age	3	47.311	0.78	0.506
Ed. & Income	12	263.931	1.09	0.367
Ed. & Age	9	81.083	0.45	0.911
Income & Age	12	211.718	0.87	0.576

* Significant at the 0.01 level

** Significant at the 0.05 level.

6. CONCLUSION

The findings of our study indicate that, at the individual respondent level, survey data are very inaccurate in measuring the respondents' actual usage of long-distance telephone communication. Our results support the earlier conclusions of Parfitt (1967), Sudman (1964) and Wind and Lerner (1982) who analyzed this issue with respect to non-service related consumer products. We cannot report, however, that our results either support or refute those of Stanton and Tucci (1984) since the time frames, and thus the recall periods, are considerably different in the two studies.

The importance of our findings extends beyond simply confirming the results of previous studies and extending the range of product types to include the analysis of a consumer service item. Our findings identify the fact that the over-reporting that occurs in surveys varies along two important demographic dimensions: respondents' household income and sex. Respondents who report very low household income tend to under-estimate usage in survey reporting while wealthier respondents do the opposite. Furthermore, this relationship tends to increase monotonically with increases in income levels and then declines. Female respondents tend to over-estimate usage in surveys by a considerably greater magnitude relative to male respondents. Taken together the findings suggest a strong possibility for measurement problems occurring if purchase data are collected using the survey method.

ACKNOWLEDGEMENTS

We are indebted to the anonymous referees for providing valuable suggestions which have improved the clarity and precision of our work.

REFERENCES

- GROOTAERT, C. (1986). The use of multiple diaries in a household expenditure survey in Hong Kong. *Journal of the American Statistical Association*, 81, 938-944.
- HIRSHLEIFER, J. (1984). *Price Theory and Applications* (3rd.Ed.). Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- KEMSLEY, W.F.F. (1961). The Household Expenditure Enquiry of the Ministry of Labour: Variability in the 1953-54 Enquiry. *Applied Statistics*, 10, 117-135.
- KEMSLEY, W.F.F., and NICHOLSON, J.L. (1960). Some experiments in methods of conducting Family Expenditure Surveys. *Journal of the Royal Statistical Society, Series A*, 123, 307-328.
- LEWIS, H. F. (1948). A comparison of consumer responses to weekly and monthly purchase panels. *Journal of Marketing*, 12, 449-454.
- McKENZIE, J. (1983). The accuracy of telephone call data collected by diary methods. *Journal of Marketing Research*, 20, 417-427.
- NETER, J. (1970). Measurement errors in reports of consumer expenditures. *Journal of Marketing Research*, 7, 11-25.
- PARFITT, J. (1967). A comparison of purchase recall with diary panel records. *Journal of Advertising Research*, 7, 16-31.
- PEARL, R.B. (1968). Methodology of Consumer Expenditure Surveys. Technical Working Paper 27, Washington D.C.: U.S. Bureau of the Census.
- SANDAGE, C.H. (1956). Do research panels wear out? *Journal of Marketing*, 20, 397-401.

- SODOL, M.G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 59, 52-68.
- STANTON, J.L., and TUCCI, L.A. (1982). The measurement of consumption: A comparison of surveys and diaries. *Journal of Marketing Research*, 19, 274-277.
- SUDMAN, S. (1964a). On the accuracy of recording of consumer panels: I. *Journal of Marketing Research*, 1, 14-20.
- SUDMAN, S. (1964b). On the accuracy of recording of consumer panels: II. *Journal of Marketing Research*, 1, 69-83.
- SUDMAN, S., and FERBER, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research*, 11, 128-135.
- WIND, Y., and LERNER, D. (1979). On the measurement of purchase data: Surveys versus purchase diaries. *Journal of Marketing Research*, 16, 39-47.

Quality Assurance Sampling for Evaluating Health Parameters in Developing Countries

STANLEY LEMESHOW and GEORGE STROH, JR.¹

ABSTRACT

A typical goal of health workers in the developing world is to ascertain whether or not a population meets certain standards, such as the proportion vaccinated against a certain disease. Because populations tend to be large, and resources and time available for studies limited, it is usually necessary to select a sample from the population and then make estimates regarding the entire population. Depending upon the proportion of the sample individuals who were not vaccinated, a decision will be made as to whether the coverage is adequate or whether additional efforts must be initiated to improve coverage in the population. Several sampling methods are currently in use. Among these is a modified method of cluster sampling recommended by the Expanded Programme on Immunization (EPI) of the World Health Organization. More recently, quality assurance sampling (QAS), a method commonly used for inspecting manufactured products, has been proposed as a potentially useful method for continually monitoring health service programs. In this paper, the QAS method is described and an example of how this type of sampling might be used is provided.

KEY WORDS: Lot sampling; Quality assurance; Acceptance sampling; Vaccination coverage.

1. INTRODUCTION

One of the problems continually confronting managers of health service programs is the identification and application of cost-effective and practical methods to monitor and evaluate operations. In developing countries the solution to such problems is usually complicated because records are often poorly maintained, reports from dispersed health facilities are usually received late or not submitted at all, and accurate target population sizes are not available. Consequently, community-based surveys are often the only means to obtain reliable numerator (*i.e.*, number of individuals with a characteristic) and denominator (*i.e.*, number of individuals studied) data. However, such surveys can be difficult to organize and implement and are often too costly to be used to monitor program operations.

Perhaps the best example of a program in which community-based surveys have been routinely used to collect information is the Expanded Programme on Immunization (EPI) of the World Health Organization (WHO) (see Henderson and Sundaresan 1982). The EPI, from its inception, has employed a cluster sampling method designed to measure immunization coverage in young children (see Serfling and Sherman 1975 and Henderson *et al.* 1973). The particular survey methodology was kept as simple in concept and application as possible to allow program managers and supervisors, often with minimal background in sampling techniques, to organize and implement the surveys (see WHO 1979). These surveys, which have been termed "30 by 7" surveys, typically involve 30 clusters and 7 individuals studied per cluster. Indeed, the strength of the EPI survey method lies in the simplicity of the design,

¹ Stanley Lemeshow, Ph.D., is a Professor of Biostatistics and Chair of Biostatistics/Epidemiology, Division of Public Health, University of Massachusetts, Amherst, MA. George Stroh, Jr., MPH, Centers for Disease Control, Atlanta, GA.

the standardized rules for implementation, and the uncomplicated procedure for compiling and interpreting results. Discussion and criticisms of the method on theoretical grounds are available elsewhere (Lemeshow *et al.* 1985 and Lemeshow and Robinson 1985).

Recently, EPI officials have recognized several practical limitations of the survey methodology. The first concern is that the results obtained with the survey method are relatively imprecise — estimates of coverage obtained can only be expected to be within 10 percentage points of the actual level of coverage in the population sampled. In developing countries where high levels of coverage have been attained, the method is too imprecise to identify significant changes between sequential surveys, or between different strata of a population being evaluated.

The second concern about the use of the EPI surveys is that, even though they are relatively easy to implement, they are still too great an undertaking for most local managers to use to assess operations in their areas of responsibility. Consequently, it is still most common for an EPI survey to be done for the entire population of a country, or for population units of relatively large size (*e.g.*: millions). Although the results are useful for managers at higher program levels, local managers and supervisors are unable to use the results at their levels of responsibility.

EPI surveys usually measure the percentage of children in an age cohort (usually 12 to 23 months of age) that should have received the entire series of vaccines that are provided in the EPI. The third concern is that this results in measurement of operations that preceded the date of the survey by more than a year; operations may have changed considerably during that interval.

Finally, an additional objective of the EPI is to develop accurate record keeping that can be used to monitor and evaluate coverage — the surveys are the primary means of assessing the validity of records. However, with the current age groups surveyed, it is often difficult to identify the set of records that correspond to the period during which immunizations were given to the children surveyed.

In this paper, we present a method which has been proposed to continually monitor a health service program and can be used to assess whether operations are maintained at an acceptable, specified level. To do this, a particular type of stratified random sampling (Cochran 1977; Hansen *et al.* 1953; Kish 1965; Levy and Lemeshow 1980) is employed that uses very small samples obtained from operationally defined units of the population. Not only can this type of community-based sampling permit monitoring of operations within relatively small populations or small areas of operation, but the results will permit managers at virtually all levels to obtain estimates to continually evaluate program operations with sufficient precision. In areas where record systems have been developed that can be used to monitor program operations, the same sampling method can be used to validate the records and ensure that an accurate numerator and denominator are available from records. Once validated these records can then be relied upon as the major source of information for program monitoring and evaluation. The general term applied to this method of sampling, which we propose as a useful alternative to more traditional methods applied in the area of public health program evaluation, is Quality Assurance Sampling (QAS) — a term well known in the areas of engineering, manufacturing and business.

2. THE QAS METHOD

The origin of QAS is in sampling and inspecting manufactured products (Dodge and Romig (1959)) where it was developed to keep labor and other sampling costs at minimal levels. One type of QAS sampling, Lot Quality Acceptance Sampling (LQAS) is identical to stratified

sampling, but the samples are too small to provide what are usually considered acceptably narrow confidence intervals for estimates for a specific stratum (usually called a "batch" or "lot" in industry). Rather, a decision is made about the quality of a particular batch or lot based on the probability that the number of defective items in the sample is less than or equal to a specified number. The results of the samples taken from all the mutually exclusive and exhaustive batches can be combined to provide a precise overall estimate of the average quality of the total product.

The strategy and goals of QAS in the health field would be similar to those in the manufacturing field. The purchaser of goods does not want to accept a batch with more than a certain percentage (P_1) defective whereas the manufacturer wants to continually monitor production to identify products with more than an expected percentage (P_2) of defectives. It is not unusual for P_1 and P_2 to be unequal. It is not difficult to see the similarities between the objectives of a manufacturer and a health manager or supervisor. The latter "produces" immunized children rather than a manufactured item.

Generally, a lot is an "operationally useful" unit. For example, in an industrial application, if there were several machines producing the same part and three operators assigned to each machine, then "lots" could be chosen that are produced by the same machine — particularly if any variation in the parts produced is most likely to be due to machine drift as opposed to operator input.

For public health work, a manager might define "lots" as recipients of services from a single operational unit — such as a health post (HP) immunization team — over a specified period of time. The amount of time between sampling could coincide with the interval between "high incidence" seasons for immunizeable diseases, but would more likely be related to the amount of time and cost associated with the sampling than any other single consideration.

In public health work a serious error would be made if the population were judged to be adequately covered ("accept the lot") when, in fact, it is not. In order to control for this possibility, we design the procedure as a one-sided test.

The null hypothesis, illustrated at the 50% level, is

$$H_o: P \geq P_o \text{ (i.e., proportion of unvaccinated children } \geq 0.50)$$

versus

$$H_a: P < P_o \text{ (i.e., proportion of unvaccinated children } < 0.50).$$

The four-celled table presented in Figure 1 describes the consequences of the testing procedure. Because the test is set up as one-sided, and because we assume the population is not adequately covered unless we reject H_o , the type I error, i.e., accepting the lot when it is defective (false negative), is the most serious error. That is, if (using the example of immunization) a population (lot) of children is thought to have an acceptable proportion immunized when, in fact, it does not, the larger number of susceptibles in the population increases the risk of transmission of the disease. Hence, we consider the "cost" of declaring that the population is adequately vaccinated, when it is not, to be high. On the other hand, the type II error, rejection of an acceptable lot, is not as serious since the result of a false-positive decision would be to concentrate efforts on an already adequately vaccinated population.

The fundamental problem in LQAS sampling, is not so much one of simply determining sample size, but of choosing an appropriate balance between sample size and critical region. In all cases, the computation of β will depend upon the actual value of P when it is assumed to be different from P_o .

		Actual Population		
		Not adequately vaccinated	Adequately vaccinated	
Decision	Fail to reject H_0	test recognizes or is sensitive to lack of adequate coverage $1 - \alpha$ sensitivity	"Provider Risk" β false positive rate	← "reject" the lot
	Reject H_0	"Consumer Risk" α false negative rate	test recognizes adequate coverage $1 - \beta$ specificity	← "accept" the lot
		"not adequate coverage"	"adequate coverage"	

Figure 1. Consequences of Hypothesis Testing in LQAS Procedure

In practice, initially a minimal level for delivery of a service would be defined on the basis of the probable distribution of service levels across lots as well as in terms of practicality (*i.e.*, a level that could be achieved). Once this level is defined, sample size options are considered relative to the number of lots that would be misclassified with stated type I and type II errors. If the sample size were too large to be practical, there would be several options including: retaining the sampling scheme, but lengthening the time interval between sampling; choosing another critical level that would allow use of a smaller sample size; choosing another QAS sampling scheme (such as double sampling or sequential sampling) that would meet the objectives of classifying the lots and still be operationally feasible; and abandoning a QAS scheme.

One means of computing probabilities and determining necessary sample sizes can be accomplished using the binomial distribution. We will assume, as is usually the case, that N is very large relative to n ; with large N , the Poisson can be practically substituted for the binomial. However, if it happens that N is not large relative to n , then the hypergeometric distribution can be used as described in Brownlee (1965) (Sec. 3.15). Letting p denote the probability of observing the characteristic, then the chance of observing exactly d individuals with the characteristic in a sample of size n is given by

$$p(d) = \binom{n}{d} P^d (1 - P)^{n - d}.$$

Suppose we decide that 7 is the sample size we wish to use. The rejection region for the test states that we should reject H_0 (and "accept the lot" as adequately vaccinated) if $d \leq d^*$. To determine the value of d^* such that $Pr(d \leq d^*) = \alpha$, we must compute $Pr(d \leq d^*)$ for a number of values of d^* . Clearly if we decide to use $d^* = 1$ then $Pr(d \leq d^*)$ would equal 0.0625 and the power of the test, if 70% of the population is actually unvaccinated, would equal 0.0038.

Results of a particular choice of n and d^* may be graphed as an **operating characteristic (OC) curve** where the variable on the horizontal axis is the proportion, P , in the population who have not been vaccinated. The vertical axis presents the probability of rejecting the null hypothesis H_0 : $P = P_0$ and concluding that the vaccination coverage in the population is adequate. Each combination of n and d^* will generate a unique curve. Figure 2 presents a typical OC curve for $n = 7$, $d^* = 1$.

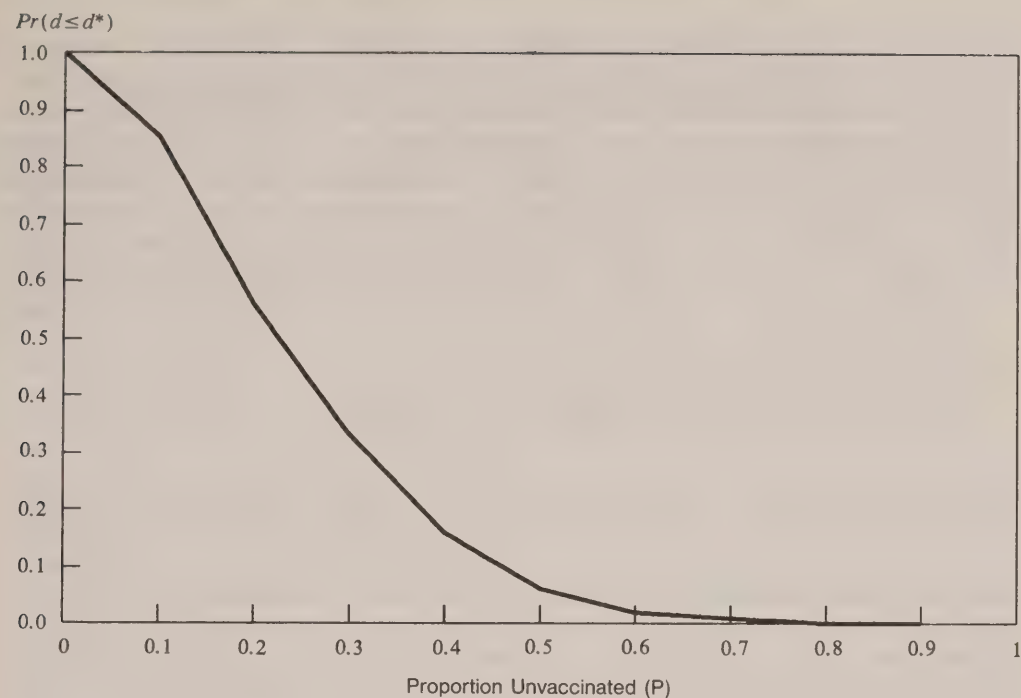


Figure 2: Operating Characteristic Curve for $n=7$ and $d^*=1$

The investigator will usually choose the value of d^* which yields a type I error less than α . Sometimes this strategy results in an extremely conservative test. For example, with $n = 7$, $d^* = 0$ and $P_o = 0.5$, α would equal 0.0078. Here the use of $d^* = 1$ with $\alpha = 0.0625$ as in Figure 2 might be justified. Table 1 presents values of d^* for small n (≤ 20) such that α will not exceed the stated type I error probability (0.01, 0.05 or 0.10) for various combinations of n and P_o . Details for the construction of this table are presented elsewhere (Dodge and Romig 1959).

The choice of the sampling scheme comes down to one of combining the desired power, $1 - \beta$, with the desired α level. Rather than providing curves which are difficult to read precisely, we developed Table 2 which presents values of (n, d^*) pairs for $\alpha = 0.05$, $\beta = 0.20$, and selected values of P under the null hypothesis (P_o) and P under the alternative hypothesis (P_a). In this table, (n, d^*) are chosen so that $Pr(d \leq d^* | n, P_o) \leq \alpha$ and $Pr(d \leq d^* + 1 | n, P_o) > \alpha$. More details are provided elsewhere (Lemeshow *et al.* 1987).

This table clearly shows the trade off one must make between power and sample size in LQAS surveys. For instance, it is essentially impossible to have $\alpha=0.05$, $\beta=0.20$ and use $n=5$ unless P_a under the alternative was actually close to 0. Hence investigators with limited resources must be ready to compromise on the value of β or the difference between P_o and P_a .

The method of quality assurance sampling described to this point is known as “single sampling” since only one sample is taken before a decision is reached regarding the disposition of the lot. A modification of this LQAS procedure, which may be useful under certain field conditions, incorporates a “double sampling” strategy. With this method, a sample is first selected of size n_1 . If this sample fails, a second sample of size n_2 may be selected. This requires the specification of two acceptance numbers. The first, d_1 , applies to the observed number of defectives in the first sample alone and the second, d_2 , applies to the total number of defectives in the first and second samples combined. In practice, the principal advantage

Table 1
Values of d^* for Combinations of P_o and n to Achieve $\alpha \leq 0.01, 0.05$, or 0.10

n	$P_o, \alpha \leq 0.01$					$P_o, \alpha \leq 0.05$					$P_o, LPH \leq 0.10$				
	0.50	0.60	0.70	0.80	0.90	0.50	0.60	0.70	0.80	0.90	0.50	0.60	0.70	0.80	0.90
5	×	×	0	1	2	0	0	1	1	2	0	1	1	2	3
6	×	0	0	1	2	0	1	1	2	3	0	1	2	3	3
7	0	0	1	2	3	0	1	2	3	4	1	2	2	3	4
8	0	1	1	2	4	1	2	2	3	5	1	2	3	4	5
9	0	1	2	3	5	1	2	3	3	5	2	3	4	5	6
10	0	1	2	4	5	1	2	4	5	6	2	3	4	5	6
11	1	2	3	4	6	2	3	4	5	7	2	4	5	6	8
12	1	2	4	5	7	2	3	5	6	8	3	4	5	7	8
13	1	3	4	6	8	3	4	5	7	9	3	5	6	8	9
14	2	3	5	6	9	3	4	6	8	10	4	5	7	8	10
15	2	4	5	7	9	3	5	7	8	10	4	6	7	9	11
16	2	4	6	8	10	4	5	7	9	11	4	6	8	10	12
17	3	4	6	8	11	4	6	8	10	12	5	7	8	10	13
18	3	5	7	9	12	5	6	8	10	13	5	7	9	11	14
19	4	5	7	10	13	5	7	9	11	14	6	8	10	12	14
20	4	6	8	11	13	5	7	10	12	15	6	8	10	13	15

×

 No test for this sample size.

Table 2
Sample Size and Decision Rule for LQAS, $\alpha = 0.05$, $\beta = 0.20$,
One-sided Test

P_a	P_o				
	0.50	0.60	0.70	0.80	0.90
	n, d^*	n, d^*	n, d^*	n, d^*	n, d^*
0.05	5, 0	×	×	×	×
0.10	8, 1	5, 0	×	×	×
0.15	11, 2	7, 1	×	×	×
0.20	15, 3	9, 2	5, 1	×	×
0.25	23, 7	12, 3	7, 2	×	×
0.30	37, 13	16, 5	9, 3	5, 1	×
0.35	67, 26	24, 10	11, 4	6, 2	×
0.40	153, 66	38, 17	16, 7	8, 3	×
0.45	617, 288	67, 33	23, 12	10, 5	5, 2
0.50		151, 80	35, 20	13, 7	6, 3
0.55		601, 340	62, 37	19, 11	7, 4
0.60			137, 86	29, 19	10, 6
0.65			535, 356	50, 35	13, 9
0.70				109, 80	20, 15
0.75				419, 321	33, 27
0.80					69, 58
0.85					253, 219

×

 Sample size less than 5.

of double sampling is that, if the defective rate is relatively low, it may be possible to study fewer subjects than with single sampling since n_1 is typically less than the n required in single sampling. However, if it becomes necessary to go to the second sample in many of the lots, the procedure may require a larger overall sample size. In most cases, the total sample size would be less than $n_1 + n_2$ since sampling stops as soon as the critical value, d_2 , is exceeded in the second sample. (The first sample is always completed to provide the information to be combined and used to compute the overall proportion acceptable in the population). Details for this procedure are presented elsewhere (Dodge and Romig 1959) and an example will be presented in Section IV.

3. ESTIMATING THE OVERALL POPULATION PROPORTION WITH QAS SAMPLING

In addition to the binary decision to "accept" or "reject" the lot, the simple random samples within each HP may be considered a stratified sample and an overall population estimate constructed.

For example, suppose 294 HP's of known population size were sampled selecting 7 children from each. Using standard stratified sampling formulae, estimates may be obtained for P , $\text{Var}(\hat{P})$, and an appropriate confidence interval may be constructed. LQAS resembles stratified sampling in that it requires that an accurate sampling frame be established in each lot and that a **simple random sample** be selected from each of these lots. However, it does not provide more information than conventional stratified random sampling since confidence intervals could be established for each stratum (or lot) and decisions could be based on values covered by each such interval (if sample sizes were made large enough to provide useful confidence intervals).

Although the n for each stratum in LQAS are too small to provide useful confidence intervals for estimates for each stratum, an appropriately designed LQAS scheme may provide a means for continually testing strata and classifying them as "acceptable" or "unacceptable" in terms of a particular outcome. This results from the fact that LQAS sample sizes are relatively small, increasing the likelihood that sampling can be done more frequently. Among its benefits, the rules of LQAS sampling are simple to follow, requiring minimal retraining of the surveyor/classifier. Lastly, since LQAS samples are, in fact, stratified random samples, the results for strata can be combined to provide adequately precise estimates for groups of strata, such as for districts, regions, or a nation as a whole.

The potential benefits of use of an LQAS scheme must be weighed against the loss of precision expected with the small samples taken in each stratum. Perhaps the best way for the reader to judge whether LQAS might be useful is an example in which a conventional stratified random sample survey approach is compared with an LQAS scheme.

4. AN EXAMPLE OF THE APPLICATION OF QAS

The example is set in circumstances similar to those in Costa Rica, and is applied to immunization coverage of children which is provided by 294 HP that cover the population of the country. The manager of the EPI would like to know the percentage of children, 12-23 months of age that received all of the immunizations that should have been given during their first year of life. Based on the immunizations that have been reported by staff, the manager thinks that the coverage level for the nation is about 60%, but the coverage that has been reported by the 294 individual HP varies from 20% to 100%; it is thought that the distribution

of coverage rates is uniform across the range. The EPI manager suspects that the estimates of coverage provided on reports may not be completely accurate because of numerator and denominator errors. As a result, it is decided that a survey of HP areas should be made in order to obtain estimates of coverage for each of the 294 areas since it would be important to be able to concentrate supervision on those HPs that have "low" coverage.

The first plan for the survey that the EPI manager evaluates is a "conventional" stratified random sampling scheme. Coverage estimates are required for each of the 294 HP, and each estimate should have confidence bounds no larger than an absolute 10%, with $\alpha = 0.05$. Since the average HP population is approximately 2500, and since it can be estimated that 3.5% of the population are children between the ages of 12 and 23 months, it is estimated that the number of children available for sampling in each HP will be approximately $2500 \times 0.035 = 88$. The formula for sample size determination which incorporates a finite population correction is given by Cochran (1977, p.75) and results in $n = 47$.

Thus, in each of the 294 HP areas, 47 (53%) of the 88 children between the ages of 12 and 23 months will be surveyed. In the entire country, 13,818 children in this age group will be surveyed. For the national estimate of coverage, P can be estimated to within 0.5% (assuming the worst level of coverage for precision (50%) and little variation in HP populations).

The manager then considers a QAS scheme. It is decided that any HP that has a coverage level of 70% or lower is performing poorly, and should be identified for increased supervision. The manager wants to be able to identify a HP with coverage of 70% with a probability of about 0.95, and HPs with lower levels of coverage with even higher probability. Several QAS schemes are considered and a double sampling scheme is proposed.

The particular double sampling scheme proposed can be denoted as $n_1:d_1 = 10:0$ and $n_2:d_2 = 14:3$. This means that in each HP area an initial sample of 10 children will be surveyed for their immunization status. Regardless of how many children are found unimmunized, all 10 will be surveyed. The number of children found unimmunized among each HP sample of 10 children will be used to compute estimates for combined areas and ultimately for the national estimate of coverage. If upon completion of a survey of the first sample of 10 children, none are found unimmunized, the HP will be categorized as having "acceptable" coverage. If 4 or more children are found unimmunized, the HP will be classified as having "unacceptable" coverage. In either scenario, no further sampling is required in the HP area. However, if upon completing the initial survey, 1, 2, or 3 children are found unimmunized, a second sample of 14 additional children is drawn. During the survey of the second sample, whenever a total of 4 unimmunized children is reached (including those from the first sample of 10) the survey is stopped, and the HP area is classified as having "unacceptable" coverage. However, if upon completion of the second sample, a total of 3 or fewer unimmunized children have been found, the HP area is classified as "acceptable".

Figure 3 shows the operating characteristic curve for this particular sampling scheme. This curve allows one to predict what the probabilities are for correctly classifying HP areas on the basis of the level of coverage. We will assume that the distribution of the 294 HPs is uniform and that all HPs in each decile have a coverage that corresponds to the mid-point value for each decile. If the probabilities of accepting a HP as having acceptable coverage are read from the OC curve and are applied to the numbers of HPs in corresponding deciles, it is possible to predict the number of HPs that would be accepted and rejected as having acceptable levels of coverage. The results of this projection are shown in Table 3.

As can be quickly computed from the expected results shown in the table, greater than 99% (183 of 184) of the HPs that had coverage less than 70% would be "rejected" (*i.e.*, they are classified as having an unacceptable level of coverage). Of the 110 HPs that had coverage above

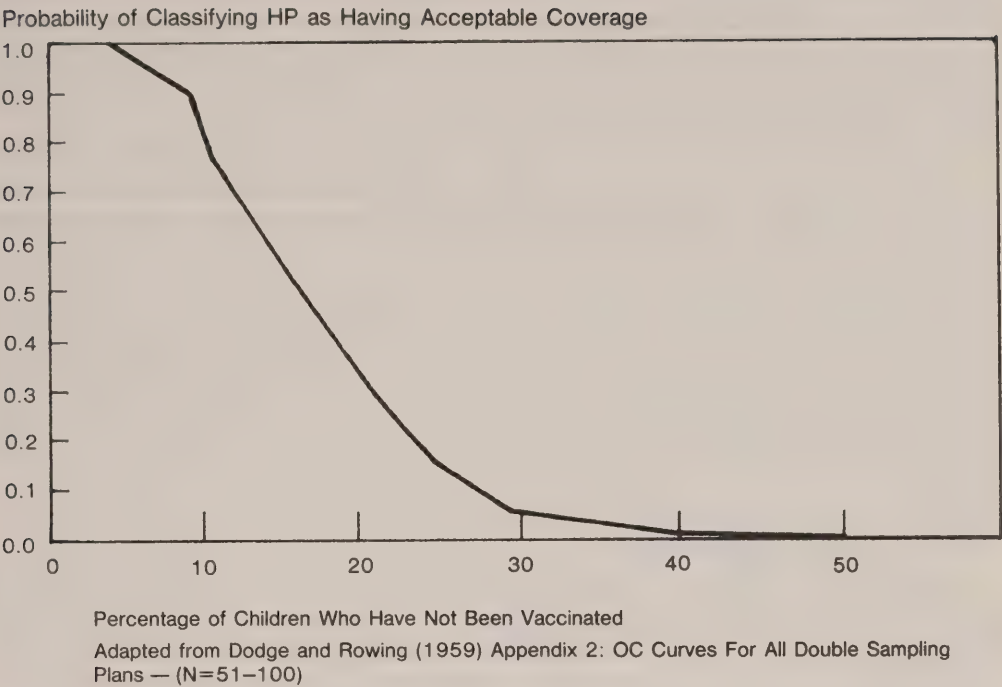


Figure 3: Operating Characteristic Curve for Double Sampling Scheme with $n_1:d_1 = 10:0$ and $n_2:d_2 = 14:3$

Table 3
Expected Classification of 294 HP with Use of Double Sampling Scheme $n_1:d_1 = 10:0$ and $n_2:d_2 = 14:3$

Percentage Coverage in HP Area	Number of HP	Number of HP Classified as:	
		> 70% Coverage	≤ 70% Coverage
20- 30%	36	0	36
31- 40%	37	0	37
41- 50%	37	0	37
51- 60%	37	0	37
61- 70%	37	1	36
71- 80%	37	7	30
81- 90%	37	21	16
91-100%	36	34	2
Total	294	63	231

Number of HP with Coverage ≤ 70% = 184.
Number Correctly Classified = 183 (99%).
Number of HP with Coverage > 70% = 110.
Number Correctly Classified = 62 (56%).

70%, 62 (56%) would be accepted (*i.e.*, they are classified correctly as having an acceptable level of coverage). Although a substantial portion of the HPs (48 of 110) that had coverage higher than 70% would be incorrectly classified as having “low” coverage, it should be noted that 63% (30 HPs) of them had coverage that was in the “marginal” range (*i.e.*, coverage levels in the 70-80% range).

Based on the initial samples of 10 children completed for each of the 294 HPs, a national estimate can be computed as with any stratified random sample. Using the same assumptions as were made for the “conventional” plan, the 95% CI for the national estimate of coverage from the QAS scheme would estimate P to within 1.8%, a level of precision that is adequate for the purpose of the EPI manager.

It should also be noted that the total number of children that would be surveyed in each HP area would vary between 10 and 24. In fact, with the particular distribution of coverage levels assumed in this example, the majority of HPs would be classified on the basis of the initial sample of 10 children (*i.e.*, of the 184 HP with < 70% coverage, about 98% would be classified as unacceptable from the initial $n_1:d_1 = 10:0$ sample). Of the minority of HPs which were not classifiable on the basis of the initial sample, few would require surveying all 14 children in n_2 . Thus, the “average” number of children sampled across all 294 HP would be substantially less than $n_1 + n_2$.

In conclusion, LQAS may have useful application in certain settings in which conventional stratified random sampling — requiring sufficient sized samples from each stratum to produce useful confidence intervals for the estimates obtained — is too costly and/or time consuming. LQAS is, in fact, nothing more than another way of interpreting data obtained with a stratified random sample with samples too small to provide meaningful confidence intervals. Because it may be possible to do such small sampling more frequently, the potential exists for establishing a system for continual monitoring of an activity, perhaps using staff that with minimal training could include monitoring activity with other field duties. One further advantage of the more frequent sampling could be that rather than concentrate on an age cohort that has passed through the full period of exposure to all immunizations, managers could instruct surveyors to collect information on children in the process of being immunized — *i.e.*, determine whether children have received the immunizations that are appropriate for their age. This would provide a means of obtaining information on more current activity, and afford an opportunity to intervene in a more timely manner to improve coverage.

Although confidence intervals will always provide much more information than a simple binary decision, the sample sizes required to obtain any useful level of precision on estimates for relatively small strata may be prohibitive. In such instances, an appropriate QAS scheme may be an alternative approach worthy of consideration.

REFERENCES

- BROWNLEE, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*, (2nd ed.). New York: John Wiley and Sons.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd. ed.). New York: John Wiley and Sons.
- DODGE, H.F., and ROMIG, H.G. (1959). *Sampling Inspection Tables*, (2nd. ed.). New York: John Wiley and Sons.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vols. 1 and 2. New York: John Wiley and Sons.

- HENDERSON, R.H., *et al.* (1973). Assessment of Vaccination Coverage, Vaccination Scar Rates, and Smallpox Scarring in Five Areas of West Africa. *Bulletin of the World Health Organization*, 48: 183-194.
- HENDERSON, R.H., and SUNDARESAN, T. (1982). Cluster Sampling to Assess Immunization Coverage: A Review of Experience with a Simplified Sampling Method. *Bulletin of the World Health Organization*, 60: 253-260.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- LEMESHOW, S., *et al.* (1985). A Computer Simulation of the EPI Survey Strategy. *International Journal of Epidemiology*, 14, 3: 473-481.
- LEMESHOW, S., and ROBINSON, D. (1985). Surveys To Measure Programme Coverage and Impact: A Review of the Methodology Used by the Expanded Programme on Immunization. *World Health Statistics Quarterly*, 38, 1.
- LEMESHOW, S., HOSMER, D., and KLAR, J. (1987). *Sample Size Determination*. To be published by World Health Organization.
- LEVY, P.S., and LEMESHOW, S. (1980). *Sampling for Health Professionals*. Lifetime Learning Publications, New York: Van Nostrand Reinhold.
- SERFLING, R.E., and SHERMAN, I.L. (1975). *Attribute Sampling Methods*. Washington, D.C., US Department of Health and Human Services, Public Health Service, Publication No. 1230.
- WORLD HEALTH ORGANIZATION. (1979). *Training for Mid-Level Managers. Evaluate Vaccination Coverage*. Expanded Programme on Immunization in Cooperation with US Department of Health and Human Services, Public Health Service, Center for Disease Control. Geneva: WHO.

European Experience of Using Administrative Data for Censuses of Population: The Policy Issues That Must be Addressed

PHILIP REDFERN¹

ABSTRACT

The experience of the four Nordic countries illustrates the advantages and disadvantages of a register-based census of population and points to ways in which the disadvantages can be contained. Other countries see major obstacles to a register-based census: the lack of data systems of the kind and quality needed; and public concern about privacy and the power of the State. These issues go far beyond statistics; they concern policy and administration. The paper looks at the situation in two countries, the United Kingdom and Australia. In the United Kingdom past initiatives aimed at population registration in peacetime foundered and the present environment is hostile to any new initiative. But the government is going ahead with a controversial reform of local taxation that involves setting up new registers. In Australia the government tabled a Bill to introduce identity cards and an associated register, and advanced clearcut political arguments to support it; the Bill was later withdrawn. The paper concludes that the issues involved in reforming data systems deserve to be fully discussed and gives reasons why statisticians should take a leading part in the debate.

KEY WORDS: Census of population; Identity cards; Personal reference numbers; Population registers; Record linkage.

1. INTRODUCTION

This paper has its origin in a study of alternative approaches to the census of population that I carried out for the Statistical Office of the European Communities (Redfern 1987). The study examined the experiences of the 12 member countries of the EEC together with Canada, Sweden and the United States. The study found that sample surveys can complement, but cannot replace, a 100 per cent census, because they do not provide reliable statistics for small areas. An important example of samples complementing a 100 per cent enumeration is the short form/long form censuses of Canada and the U.S. A sample survey complementing 100 per cent data from registers is in prospect in Norway (Section 3.3).

Registers that contain addresses give figures for small areas; and, if the registers cover the census topics reliably (in terms of definitions, coverage, accuracy and timeliness) and can be linked, it is possible to create a record for each individual akin to his census return and so to conduct a register-based census: in essence administrative data are being recycled for statistical purposes. The pressure of costs and the burden of formfilling in the traditional census have persuaded the Nordic countries (Denmark, Finland, Norway and Sweden) to adopt this approach in whole or in part.

Though administrative data can support a *conventional* census in a variety of ways (Redfern 1987, paragraphs 3.65—3.67), it is their use in a *register-based* census that provides the first main theme of this paper. Section 2 describes the registers that are needed as a base for a census and Section 3 identifies the similarities and differences between the four Nordic countries in

¹ Philip Redfern, 17 Fulwith Close, Harrogate, North Yorkshire, England, HG2 8HP.

their approaches to this kind of census. Section 4 then considers the obstacles that other countries would face if they were to upgrade their record systems so as to make a register-based census feasible, and recognises that the issues raised concern administration and policy more than statistics.

It is these wider issues that provide the second main theme of the paper. Section 5 looks in more detail at a country in which, for reasons of policy and ideology, administrative records are not coordinated through a population register: the United Kingdom. Section 6 describes a recent initiative in Australia to improve administrative records. Finally Section 7 summarises the political arguments for and against coordinating administrative records through population registers and puts the case for statisticians taking a leading part in debate on the subject.

2. THE REGISTERS NEEDED AS A BASE FOR THE CENSUS

2.1 Population Registers

The essential starting point for a register-based census is a population register that includes personal reference numbers and addresses. The personal numbers must be in one to one correspondence with the members of the population. To keep the register up-to-date the citizen is obliged to notify changes. The personal numbers are also recorded in the files of the various administrative agencies, and so can be used to link records for statistical purposes.

Population registration serves essentially administrative ends. It is an efficient way of organising the many dealings between public authorities, both central and local, and the individual citizen: for example taxes, social security, publicly-provided health services and electoral registration. To work effectively, population registration should serve a wide range of administrative activities, so that opportunities for updating and correction are frequent and the citizen becomes used to quoting his personal number.

The key to the system is the central population register which records identifying information about each person (name, place and date of birth, date of immigration, marital status, and possibly items like parentage and citizenship) and his permanent reference number. In most countries the central population register includes up-to-date addresses, though the French *Répertoire National d'Identification des Personnes Physiques* does not. The basic administrative function of the central register is to act as reference point for administrative agencies which can check the identities of the individuals that they are dealing with and, as necessary, can correct or record the personal reference numbers in their own files.

2.2 Other Key Registers in a Register-Based Census

A register-based census of population and housing makes use of registers of other kinds of units than persons. The most important are a central register of housing and a central register of business enterprises and establishments (workplaces). Provided the housing register identifies each housing unit (and not just the building or the address) with a code that also appears as part of the address in the population register, then data on the housing unit in the housing register can be associated with data on the occupants in the population register: the two registers can be linked. Similarly a register recording each person's employer and workplace can be linked to a central register of enterprises and establishments to show the person's industry, commuting journey, *etc.*

3. CENSUSES IN THE NORDIC COUNTRIES

The four Nordic countries have well-developed population registers of the kind described in section 2.1. They have constructed, or propose to construct, central registers of building and housing to serve mainly administrative purposes. This section of the paper outlines the census of each country in turn and then summarises the directions in which Nordic census-taking is developing.

3.1 Denmark

Denmark is the only Nordic country — and I believe the only European country — to have switched completely from the conventional census to a register-based census. The switch was made in little more than a decade. The central population register with personal reference numbers was created in 1968 for administrative purposes, and a register-based census of population (but not housing) followed in 1976. A central register of buildings and dwellings was created in 1977, again mainly for administrative purposes, and a register-based census of population *and* housing followed in 1981. Another significant step in 1979-80 was to extend the return in which employers report each employee's earnings to the tax authorities: employers with more than one workplace added each employee's workplace to the return. This was done purely for statistical purposes and the statistical office has had to make a considerable effort to secure a good response.

The registers held by Danmarks Statistik for statistical purposes, numbering some 37, provide annual or more frequent statistics of population, employment, commuting, income, housing and construction for municipalities and sometimes smaller areas. But, because of the cost, analysis on the scale of a census takes place much less frequently: the next after the 1981 census will take place in 1991 and even that may be on a lesser scale than 1981.

The transition to a register-based census has been facilitated by the reorganisation of the Danish central statistical office in 1966. Danmarks Statistik was given a measure of independence of the central government, which could help to reassure the public on confidentiality. It was given powers to demand, and to use for statistical purposes, data held by public authorities for administrative purposes, and to participate in the construction of registers containing such data.

The problems that Danmarks Statistik now faces concern mainly the quality and timeliness of data, both of which depend on the efficiency of administrative procedures. Thus the slowness in compiling tax authorities' files — which provide data on industry, occupation, journey to work and income — delayed analysis of these topics in the 1981 census until summer 1983; and it is expected that statistics on the labour force will continue to lag at least a year behind the reference year to which they relate. Reliable data on occupation are particularly difficult to obtain because the topic is of little administrative interest; a main source is the information given by the taxpayer on his annual tax return. Despite problems of these kinds Danmarks Statistik takes the view that the register-based census has come to stay in Denmark because of the savings in cost and in burden on the public (Jensen 1983).

3.2 Finland

Register-based censuses have a long history in Finland. In the 1600s the parish registers recorded everyone over the age of 12 living in the parish, and in 1749 figures of the total population were compiled analysed by age, sex, marital status and social class: one of the first-ever register-based censuses? Later censuses followed this pattern. The censuses of 1950 and 1960 adopted the conventional method of collecting the information through questionnaires. But

beginning with the 1970 census an increasing range of data has been extracted from registers. In the mid-decade census of 1985 the questionnaire asked only about economic topics: type of activity (if any) and occupational status, employer and workplace, occupation, and number of months worked in the past year. Data on housing were taken from the register of buildings and dwellings that had been created from 1980 census data and is updated with information from the municipalities.

The 1985 census was planned to cost a little under the equivalent of 1 US dollar per person, or only a quarter of the cost of the 1980 census in real terms though covering the same range of variables. Factors that helped to make this possible included: mail-out of questionnaires preprinted with data on workplace (from the 1980 census) and occupation (from the central population register) — to be corrected by the respondent if necessary; mail-back to the central office with no local field organisation; only one reminder, with no follow-up of the 3.7 per cent of forms which were not mailed back or were mailed back incomplete; and imputation of missing data, where possible, using a variety of registers, one of which was pension records in respect of private sector employment. The final response rate to the questionnaire was 97.4 per cent, and by imputing missing data a final coverage of 98.6 per cent was achieved. Another reason for the low cost of the census is that part of the cost and burden has been transferred to the registration systems, including the annual field checks on the population registers by means of forms issued to each household/dwelling and quinquennial checks on the register of buildings and dwellings by means of forms sent to owners and occupiers.

Comparisons between the 1980 census responses and register data on economic variables have been regarded as encouraging. This, and the methods developed in the 1985 census to impute the economic characteristics of non-respondents, open up the possibility that the 1990 Finnish census might be wholly register-based. To fill one gap in register data, employers with more than one workplace will in future make a return of each employee's workplace (Laihonon and Myrskylä 1987; Heinonen and Laihonon 1987).

3.3 Norway

The 1980 census of Norway was to a substantial extent register-based. It took data on basic demographic topics, income and completed education (other than education abroad) from registers. These data were complemented by means of a mail-out mail-back questionnaire to each person aged 16 and over on economic topics, education abroad, country of birth, religious affiliation and housing. All persons in the same household were to return their forms, together with one housing form, in the same envelope, thus defining the composition of the household for census purposes.

For several reasons it is not feasible to switch to an entirely register-based census in 1990. First, register data on some important census variables do not conform to desirable statistical definitions or are not of sufficient quality for census purposes (this applies for example to industry); and register data for other variables do not exist (for example occupation). Second, the development of the register of land property, addresses and buildings (the "GAB" register), begun in 1983, is unlikely to be far enough advanced by 1990 to provide housing data for the census. Third, because the link between the GAB register and the population registers is the address, it is not possible to identify household composition or to associate housing characteristics with personal characteristics when two or more housing units have the same address.

In the 1990 census data from registers will again be used for basic demographic topics, income and completed education (other than education abroad). A method is being developed for converting register data on most of the economic variables to statistically-desirable definitions by

reference to the results of an enquiry addressed to a 10 per cent sample of persons aged 16 and above (100 per cent in municipalities with populations under 6,000). The register data for a sub-population would be adjusted in part using sample data for the sub-population and in part using sample data for a wider population — a procedure that would partially eliminate the bias in the register data. The sample enquiry would be the only census source for topics for which no register data exist, including occupation and probably housing and household composition.

This approach — the use of registers plus a 10 per cent sample enquiry — is estimated to cost 60 per cent of the cost of a census on 1980 lines. The penalties would be the sampling variance, which would be greatest for topics for which no register data exist, and also some bias in the case of topics for which register data exist but are not of the quality needed for census purposes (Heldal *et al.* 1987).

3.4 Sweden

Over the past two decades the balance of the Swedish census has changed: in 1970 most of the data came from questionnaires and a few from registers, but in 1985 the position was reversed. In 1985 the mail-out mail-back questionnaire to each person aged 16 and over (or married couple) asked only (1) whether the person was economically active in a specified week and, if so, the occupation, (2) the household composition — a list of the adults who live in the dwelling and (3) housing questions. It was possible to omit questions asked in the preceding census on the name of the enterprise at which the person was employed, the workplace and the industry, because from 1985 the annual returns that employers make to the tax authorities giving each employee's earnings were extended to show the employee's workplace. But the topic hours of work was dropped from the 1985 census when employers resisted the proposal to include this too on the annual returns.

After the 1980 census a study had been made of the steps that would have to be taken if the 1985 census were to be wholly register-based. The steps included:

- (1) The use of data on occupation from the forms on which employed persons report changes in income to the national insurance offices.
- (2) The creation of a register of household composition, which would be updated by asking for more information when a person moved house.
- (3) The creation of a register of buildings that contain housing units, to be updated by the municipalities.
- (4) The creation of a register of completed education, to be updated with information from educational institutions on new graduations.

But, as already noted, a questionnaire was retained in the 1985 census mainly because of doubts about the quality of information that could be obtained from registers on occupation, household composition and housing. Of the proposed new registers only the register of completed education is as yet under construction. But a committee is studying the possibility that the record of a person's address in the population registers should include the housing unit and not just the property — an essential step in linking population registers to housing registers.

A Parliamentary Commission is reviewing the 1985 census, particularly aspects concerning privacy and confidentiality. Its findings will be one of the factors shaping the 1990 census.

3.5 Summary of Nordic Census-Taking

The four Nordic countries are developing their censuses along different paths but there are many features in common:

- (1) All have as a starting point accurate registers of population which give regular and reliable statistics of population for small areas.
- (2) All wish to maximise the use of information in other registers and to minimise the burden of formfilling on the public. All are striving to contain or reduce costs.
- (3) All recognise the problems of definition, quality and timeliness of the information in registers, particularly for economic topics. Employers' returns are being extended to give information on each person's workplace, and hence on industry — though extensions for purely statistical purposes are unwelcome and may yield data that are of poor quality. Register data on occupation are generally unreliable. And data on some topics, such as method of travel to work, do not exist in any register.
- (4) Registers of buildings and houses have been created or are proposed. But it is difficult to keep the registers up-to-date, whether by using information available to the municipalities or by collecting information directly from owners. In some countries the registers need to be further refined to identify each housing unit in a way that permits a link with the address information in the population registers. Another problem is how to get data on household composition from registers if, as in Sweden, the household is not defined as all the occupants of the housing unit.

All four countries appear ready to sacrifice something in the quality of the census results in order to cut costs and the burden on the public. But they differ in their approaches. Denmark has gone the farthest by abandoning the census questionnaire. Because of doubts on the quality of some register data, particularly on economic topics, the 1985 censuses in Finland and Sweden retained a limited questionnaire, and the responses were linked to demographic and other data taken from registers. But the possibility is foreseen of making the 1990 census of Finland wholly register-based. In Norway, where there was no mid-decade census, the 1990 census is expected to retain a questionnaire on at least economic topics but, to reduce costs, the questionnaire may be sent only to a 10 per cent sample of persons; where register data for economic topics exist, though imperfect, they could be converted to statistically-desirable definitions by reference to the sample data. A valuable account of Swedish experience of using registers as a census source has been given by Johansson (1987).

4. THE FEASIBILITY OF A REGISTER-BASED CENSUS IN OTHER COUNTRIES

The two main forces that have driven the Nordic countries towards a register-based census — the need to cut costs and the burden of formfilling — have been strongly at work elsewhere. They show for example in a halt, and sometimes a reversal, of the pre-1980 trend to longer census questionnaires.

A new and disturbing feature, public protest, disrupted the census in two countries. In the Netherlands the plans for a 1981 census were abandoned. The census in the Federal Republic of Germany planned for 1983 had to be postponed to 1987 because of more stringent conditions on confidentiality laid down by the Constitutional Court, and even then there was some non-cooperation. No country can feel itself secure against this kind of challenge. But a register-based census is less likely to be sabotaged provided it does not have to be supplemented by a questionnaire. This is because there is no occasion (Census Day) when everyone is faced with a questionnaire and the protests of a minority can be fanned into large-scale opposition.

If the register-based census is so much cheaper with less burden on the public and less risk of sabotage, why do so few countries see it as a viable methodology? There are three main reasons. First, for some topics, particularly economic topics, administrative data may be of poorer quality than data collected through questionnaires; and for other topics no administrative data exist. The Nordic countries recognise these shortcomings, and so some have retained a questionnaire and linked the responses to the data from registers (Section 3.5).

Second, many countries do not possess the necessary data systems of the kind described in Section 2. For example, local population registers may exist but without a central population register, as in the Federal Republic of Germany, Greece and Italy. The population registers may not be up-to-date and indeed some countries rely heavily on the canvass for a conventional census of population to update the registers (Italy and Spain). Outside the Nordic group, the Benelux countries have, or are likely soon to have, the data infrastructure needed for a register-based census.

The third main obstacle to a register-based census follows from the second. If the data systems have to be radically improved — and particularly if there has to be wider use of personal numbers and a new obligation to notify each change of address — opposition may be expected from politicians and the public on grounds of privacy and erosion of freedom. There may be doubts too whether the public would cooperate in the bureaucratic disciplines of a good register system. In addition, even when the necessary data infrastructure is in place, its use for record linkage for census or other statistical purposes could be sensitive. These are important issues but they go far beyond statistics. They concern policy and administration. They are now discussed by reference to the experience of the United Kingdom.

5. RECORD SYSTEMS IN THE UNITED KINGDOM

Decennial censuses in the United Kingdom use conventional methods. The 1981 census was probably the most successful census since the Second World War — a success that was helped by the shortened form and the omission of a controversial question on ethnicity. So three factors combine to make a register-based census seem a rather remote possibility: the 1981 success; doubts about the range and quality of statistics that could be extracted from administrative records; and the absence of a population register to coordinate the record systems.

But statisticians have recognised the benefits, both administrative and statistical, that population registers could bring. The two initiatives on this subject in the past 70 years — both of which failed — are described in Sections 5.1 — 5.4. Now the government, while opposing a central population register, is introducing a limited form of local population register as part of a controversial reform of local taxation (Section 5.5).

5.1 National Registration in Two World Wars: The 1918 Committee on Registration

Thinking in Britain about population registers goes back over seventy years to the First World War. The National Registration Act of 1915 had obliged every adult to carry a National Registration Certificate and to register every change of address. This led Sir Bernard Mallet, Registrar General, to consider a permanent system, which he outlined in his Presidential address to the Royal Statistical Society in November 1916 (Mallet 1917). But he was aware that he might be criticised for “desiring to Prussianise our institutions”.

These ideas were developed in the report of a committee appointed by the government in 1918 and chaired by Sir Bernard Mallet. Many years later he reviewed the findings in his Presidential address to the Eugenics Society (Mallet 1929). What he then said remains true today:

“We found in existence in England a very considerable number of registers being kept at considerable expense for various special purposes, some of them covering very large sections of the population. These registers are kept under different Acts of Parliament, by various authorities, in varying areas, for independent purposes, without any provision for their coordination one with another”.

The committee proposed continuous registers of the population kept locally and associated with identity cards. A central index register would interrelate the local registers to deal with removals and to prevent duplicate entries. This registration system would coordinate the registers kept for special purposes — electoral registers, school attendance registers, the decennial census, registers of births, marriages and deaths, etc. It is noteworthy that the committee, reporting nearly seventy years ago, proposed that the census of population should be linked to population registration.

In his 1929 address Sir Bernard Mallet set out the principles to which any good system should conform: first, the accurate identification of every individual “in order (a) that he shall be made responsible for the fulfilment of his obligations to the community and (b) that he shall be ensured his rights as a citizen, whether these take the form of franchises to be exercised or dues to be received”; second, the acquisition of statistical information and in particular regular figures of the populations of local areas. The analysis made and the proposals that followed would still stand as a valid response to the situation that we face in the United Kingdom today, though some of the features would not be acceptable now. Thus:

“the numerous official enquiries and registers, now made and maintained independently of each other, would be coordinated into a single system which would provide a *dossier* for each individual containing those particulars regarding him which the State is concerned to know” (Mallet 1929).

To Sir Bernard Mallet’s regret the recommendations in his committee’s report were not carried out and, with the demise of the temporary wartime legislation, national registration ceased until the outbreak of the Second World War.

During the Second World War and for a few years after a full system of population registration operated in Britain. A National Register was set up linked to the issue to each person of an identity card bearing his identity number and address. Local registers were coordinated through a central register which held each person’s name, date of birth, identity number and a code for area of residence. A person had to notify changes of address to the local register. The National Register survived until 1952 when identity cards and the obligation to notify changes of address were abandoned in a post-war spirit of “set the people free”.

5.2 The National Health Service Central Register

The central register set up in 1939 during National Registration has been maintained since 1952 to serve a more limited role in the running of the National Health Service (NHS). Renamed the National Health Service Central Register (NHSCR), it now includes everyone resident in Britain apart from the 1 or 2 per cent who were born abroad and who have never registered on the patient list of a doctor in the NHS. But the NHSCR does not fill the role of a central population register of the kind found in many countries in Northern Europe because it is not used as a reference point from which other agencies can check personal identities and can carry the personal reference numbers into their own files. Indeed the identity numbers recorded in the NHSCR serve only NHS purposes. Other limitations which would inhibit the wider use of the NHSCR are:

- (1) A significant proportion of the data arriving at the NHSCR do not carry the identity number and, given the difficulty in using names and dates of birth as unique identifiers, some of these data cannot be linked to already existing NHSCR records; thus some 1 or 2 per cent of the deaths notified to NHSCR cannot be linked in. This and the failure to remove all emigrants from the register are main factors in the inflation in the register, currently estimated at about 5 per cent. But this figure should reduce shortly when the register is computerised.
- (2) Addresses are held in full in local registers and as area codes in the NHSCR. But in most cases changes of address are recorded only when a person registers with a new doctor — which may occur years after the person has moved house.

5.3 The Wide Range of Registers in the United Kingdom

As in any other developed country, a wide range of registers containing personal data is held by public authorities in the United Kingdom. The main ones concern vital registration (births, deaths, marriages and divorces), immigration and naturalization, the national health service, social security (contributors and beneficiaries such as the unemployed, pensioners and children), personal taxation, passports, electoral lists, the ownership of cars and licenses to drive cars. But these registers are maintained independently of one another by the different agencies, each with its own personal numbering system. (An exception is the joint arrangements for collecting employees' social security contributions and income tax under Pay-As-You-Earn, using one set of personal numbers, the National Insurance numbers.) This case apart, there is no coordination of record systems, no consistency in the content of records and no single set of personal numbers in general use. Details of a person's identity, usually name and date of birth, may differ between one register and another or even within the same register. This causes duplication and makes linking between registers for statistical purposes uncertain and costly. Information on address is even less consistent. There is no mechanism for carrying updating information simultaneously into all relevant records, for example information on change of address, change of name on marriage, or even the fact of death. In the words of Sir John Boreham, then head of the Government Statistical Service (GSS), "the information is never properly brought together . . . It's all rather ramshackle" (Boreham 1985).

5.4 The 1960s Study of Registers

The existing uncoordinated system of records is inefficient for administration; and the absence of up-to-date addresses and the inability to link records are severe handicaps for statistics. And so in the late 1960s the GSS looked for a remedy. It studied the case for replacing the variety of personal numbering systems by a single set of personal numbers to be held in a central register, which might also include up-to-date addresses (Penrice *et al.* 1968). But Ministers decided that these ideas were politically unacceptable and terminated the studies (House of Lords 1969).

5.5 The Registers for the New Community Charge

It would seem that one of the biggest obstacles to the creation of a population register in Britain has now been overcome: an obligation has been laid on the citizen to report changes of address. Despite this, no effective population register will be created. The government has set its face against that.

The new obligation to report changes of address — a revolutionary departure from peacetime traditions in Britain — stems from the government's decision to change the basis of local taxation. In the past local taxes have been levied on the occupiers of property on the basis of the property's rental value. The tax on the occupier of a dwelling is now to be replaced by a flat rate tax on each person aged 18 and over living in the dwelling: the *Community Charge* (CC). To administer the tax new local registers will be maintained listing addresses and the persons aged 18 and over resident there. Though the registration officer will be able to make enquiries and to call on information held by local authorities and housing bodies and in electoral rolls, the obligation to inform him of changes to the register is laid on the individual. Legislation has been enacted to introduce the new system in Scotland with effect from April 1989 and in England and Wales from April 1990.

But the CC registers will be primitive instruments compared to the population registers in the Nordic and Benelux countries because:

- (1) The CC registers will not cover everyone; in particular they will not cover the under-18s and people living in boarding houses and institutions.
- (2) The registers (which will record each person's name, address and, in Scotland only, date of birth) will be maintained locally with a limited degree of standardisation of procedures. There will be no central register to standardise the description of each person's identity and to coordinate the local registers (for example to facilitate transfers between authorities).
- (3) Although the legislation makes no specific provision for including a personal reference number in the registers, a report had recommended that local authorities in Scotland should create such a number and suggested a possible algorithm for this based on name and date of birth (Chartered Institute of Public Finance and Accountancy 1987). But the recommendation is not being implemented.
- (4) The legislation specifies who can have access to which parts of the register. Apart from local authority access for the purpose of administering the CC: an individual can inspect the entry relating to himself; the public can inspect the list of addresses and the names of persons relating to each address (but, to quote the Scottish legislation, "not so as to ascertain whether that person resides at that address"); and the Electoral Registration Officer has access for his purposes. No other access is permitted.

The government's rejection of a population register that would coordinate administrative records is spelt out in the Green Paper on the CC scheme (Her Majesty's Government 1986). The paper cites countries that "have unified their separate registers and use them for several different central administrative purposes". It goes on "The British tradition is different. Registers are kept separately for different purposes by the body which needs them for a particular purpose. . . . There will be no national register." This contrast between other countries' practices and United Kingdom practice is mistaken, because in other countries the different agencies maintain *separate* registers but call on a central register in order to identify the individuals that they are dealing with. I would judge that the statement "There will be no national register" reflects a political axiom, not the conclusion of rational analysis.

The creation of the CC registers is perhaps a missed opportunity to set up an effective population register. But the CC scheme is not an ideal vehicle for that. If it is to be effective, population registration should serve many ends, the more the better, and not just one — particularly when the single purpose is to levy a tax which many will feel onerous and many may try to avoid. Moreover the CC is politically controversial because of its differential impact on various groups in the community: in general terms a transfer of resources from the poor to the rich.

Thus there are several reasons for questioning the operational effectiveness of the registers to be set up under the CC scheme: the single purpose and controversial aim of the registers; the incomplete coverage of the population (the omission of some groups); the lack of a central register to coordinate the local registers; and the reliance on a person's name and (in Scotland only) date of birth as identifiers rather than a permanent personal number. The local authorities have made some critical observations on the problems that they will face in attempting to set up the registers (Rating and Valuation Association 1987). It looks as though the government has embarked on new tax legislation without thinking through the practicalities of implementation.

Another worrying feature of the CC scheme is its effect on response to the 1991 census of population. Many of those who evade CC will probably try to evade the census too, not trusting the census authorities' assurances that census data will not be passed on to other agencies. And if the census form is too explicit by stating "YOUR INFORMATION WILL NOT BE PASSED TO THE AGENCIES DEALING WITH TAX, SOCIAL SECURITY, COMMUNITY CHARGES, . . .", will the census authorities themselves be seen to be condoning or even encouraging evasion and fraud?

5.6 The United Kingdom Environment

Leaving aside the CC, the present environment in the United Kingdom is generally hostile to the idea of population registers. But two positive features may be mentioned. First, the *Data Protection Act, 1984* introduced safeguards for personal data held on computers on the lines of the Council of Europe's Convention of 1981 (Council of Europe 1981). In fact the government's primary aim in introducing the 1984 legislation was commercial: to establish the United Kingdom as a safe place in the eyes of other countries which might be considering transmitting their data to the United Kingdom for processing. Protection of privacy was a lesser aim. Second the GSS, which would be concerned with some aspects of the working of population registers, has established an unquestioned record of protecting data; it has published a code of practice (Government Statistical Service 1984). Integrity in handling data has been underpinned by the fact that the GSS is decentralised, so that legal and administrative barriers have prevented the exchange of data even for statistical purposes. Such barriers would have to be removed if the statistical fruits of population registration were to be secured.

On the other side of the balance sheet the GSS's dependence on central government contrasts with the relative autonomy of the statistical organisations in, for example, Denmark and the Netherlands; this could lessen public confidence in its handling of data. The GSS's image as a creature of central government has been intensified by the Rayner Reviews of the early 1980s, as a result of which the GSS was instructed to give greater priority to the needs of central government at the expense of the needs of others — the local authorities, business, academics and the general public.

A main obstacle to population registers in the United Kingdom is the public's traditional resistance to governmental actions that appear to be overbearing or bureaucratic. The privacy lobby can be relied on to lead the opposition to any new reporting obligations placed on the public, to any extensions of the government's holding of personal data or to any project for linking data. The opposition overlooks the costs and injustices that result from inefficient management of data; and it overlooks or undervalues the checks on the misuse of personal data that can be provided by legislation on data protection and freedom of information — if properly implemented. In recent years fears about giving more personal data to the government have been reinforced by the public's perception of the style of government: the United Kingdom government is seen as almost obsessively secret and as seeking to concentrate power

in its own hands. Thus, not only is there no Freedom of Information legislation in the United Kingdom, but all government information has, in principle, been protected by the catch-all *Official Secrets Act, 1911* (Superseded in May 1989 by a more narrowly worded Act). Peter Hennessy, editor of *Contemporary Record*, asserts that British governments “maintain the tightest system of administrative secrecy in the western world” (Hennessy 1987). And recent events have called into question the proper accountability of the security services. Writing of the whole range of government activity, William Plowden, Director General of the Royal Institute of Public Administration, said “a modern British government, supported by an adequate majority in the House of Commons, at little risk from the rubber-toothed bulldogs of the select committees and entrenched behind the Official Secrets Act, is one of the least accountable executives in the developed world” (Plowden 1987).

So the public is suspicious of any new scheme of population registration. And, as already noted, opposition to full registration has been expressed by the present administration, which, like its counterpart in the United States, has made determined efforts to “get government off our backs”. One of the administration’s major policy objectives has been to reduce the size and influence of the public sector – sometimes giving a higher priority to this than to cost-effectiveness. So public concern about privacy, political ideology and scarce resources combine to block a full register which could lead to substantial savings and to a fairer and more just society. In fact there has been no balanced presentation of all the issues, and so no public discussion of them, in the past half century.

6. AN AUSTRALIAN INITIATIVE: IDENTITY CARDS

I know little about the Australian temperament or the Australian political scene, but I guess that resistance to bureaucratic government is as strong there as it is in the United Kingdom. Even so, the Australian government introduced a Bill to issue each citizen with an identity card – the Australia Card (AC). The reasons were wholly administrative: to reduce tax evasion, to reduce social security fraud and to reduce illegal immigration. The AC would carry the person’s name, his photograph, his signature and an AC number (personal reference number) but not address. It would be backed up by an AC register (which would also include address and date of birth) accessible only to certain government departments.

The *Australia Card Bill, 1986* was passed by the House of Representatives but was rejected by the Senate (in which the government party did not have a majority). The rejection was given as one of the reasons for calling the July 1987 general election and, following the electoral success of the government party, the Bill was due to come before Parliament again. But the Bill was withdrawn because of a serious legal flaw. However it is worth describing the Bill’s provisions.

The AC register would be a central population register. But it would be less developed than those in Northern Europe for two main reasons:

- (1) The Bill did not place an obligation on the citizen to notify each change of address. The hope was, I understand, that most changes of address would be picked up by one or other of the government agencies taking part in the scheme and would then be passed on to the AC register.
- (2) The AC scheme would not be as multi-purpose as several of the population registers in Europe. As a result of concerns about privacy and uncontrolled linking of data, the AC register would be accessible only to the government agencies dealing with tax, social security and health insurance, and then only to check identities.

The Bill defined the situations in which a person could be required to produce his AC; these included making any of a wide range of financial transactions, entering a new employment, claiming Medicare or social security benefits, and receiving hospital treatment. It would be illegal to require a person to produce his AC in any other situation.

As a further safeguard on privacy the Bill provided for a Data Protection Agency. However the government argued that privacy had to be balanced against the losses to government funds through tax evasion and fraud. The government estimated that the costs to government of the AC scheme would \$0.8 billion over ten years, but that this would be offset many times over by savings of \$4.1 billion in tax and \$1.4 billion in social security, giving a net saving over the ten years of \$4.7 billion (Australian House of Representatives 1986).

Remarks made by the Minister of Health in Parliament (Australian House of Representative 1986) show what Ministers were trying to achieve and the clear political commitment:

“I bring before Parliament today . . . a long overdue reform to provide fairness and equity for all Australians.”

“No one doubts that the Australia Card will check tax evasion; no one doubts that it will contribute to the integrity of our social security system; no one doubts that it will be a useful weapon in deterring illegal immigration; no one doubts that by facilitating the pursuit of the money trail it will provide an invaluable instrument against corporate and organised crime.”

“Irrefutably, citizens need to be protected against abuse of their privacy by government. But equally citizens need to be protected against others who cynically hide behind the mantle of privacy to create false identities and thus defraud the community.”

“It is inevitable that this country will establish an identification system before the century is out.”

Though the AC Bill has now been withdrawn, the government is searching for other ways to clamp down on tax and social security fraud, and so the story is not yet ended.

6.1 Identity Cards

The main emphasis in the Australian scheme was placed on the identity card as a way of checking identity, rather than on the personal number and register. Some European systems also combine the issue of identity cards with population registration; the Belgian system is one of the most highly developed. And undoubtedly the identity card provides an extra tier of security — provided it is not forged or stolen. In some countries identity cards are unconnected with population registration, for example in France.

In countries unaccustomed to identity cards in peacetime, the card is seen as a symbol of an authoritarian régime and an affront to civil liberties. That may be one of the reasons why the AC scheme generated so much public opposition in Australia. But much of the benefit from population registers can be secured without identity cards provided that citizens know their personal numbers and quote them in dealings with public authorities. This is what happens in Denmark and Sweden where population registration is effective, both administratively and statistically, without issuing identity cards to everyone.

A country like the United Kingdom ought not to shy away from correcting the incoherence of its records just because the uninformed critic might equate the necessary remedy — population registration — with what is only an optional extra — identity cards.

7. CONCLUDING REMARKS

Setting up a population register, with up-to-date addresses and personal reference numbers that are also carried into administrative files, would in fact be little more than bringing order into an existing ‘‘ramshackle’’ system: even in the most ramshackle system the citizen has to identify himself and inform various agencies of a change of address. Nonetheless some people are deeply worried by the prospect of a population register because of its threat to privacy and freedom and because it gives increased power to the State with all the dangers of misuse by an authoritarian or oppressive government. But specific remedies can and should be put in place: an effective data protection régime and legislation on freedom of information.

On the other hand a properly coordinated record system would have political advantages that have been largely overlooked. At the top of the list I would put two things:

- (1) A brake on fraud, crime and illegal immigration.
- (2) A fairer society, so that burdens and duties are fairly shared and benefits and rights go only to those entitled to them. Put another way, freedom should not extend to the freedom to cheat the rest of the community.

Rather lower down the list I would put:

- (3) The financial savings to government. More accurate records will cut the costs of administration, give a higher yield of tax and reduce the amount of benefits paid improperly — illustrated by the Australian figures (Section 6).
- (4) A wider range of policy options for government. Thus, if a reliable population register were already in place in the United Kingdom, the government would not have to construct a register *ad hoc* in order to launch its Community Charge scheme; and it could regulate immigration through control on residence in addition to the controls at airports and seaports.
- (5) Other benefits from more reliable checks on identity. The late Registrar General gives as an example better checks on a couple’s eligibility to marry. There would also be fewer different reference numbers to be quoted and perhaps fewer plastic cards to be carried.
- (6) Better statistics (but see a qualification below).

This list is one answer to the charge that a population register is totalitarian and Big Brother. Without safeguards and in the wrong hands it could be. But it could also be the key to a fair and just society. The question is: what kind of society do we seek? Is it one that encourages, or at least turns a blind eye to, fraud, tax evasion and crime? Australian Ministers cite the man who was convicted for collecting over 50 separate unemployment benefit cheques each fortnight (Australian House of Representatives 1986). In the United Kingdom a Member of Parliament and barrister was convicted in 1987 for making multiple applications for shares against the rules by using different names, addresses and bank accounts; the defence was that it was common practice.

Another answer to the charge of totalitarianism is to look at the population registers in other countries. Table 1 divides 15 countries — all the countries of Western Europe except Austria and Switzerland — into four groups according to the kind of register system that each has. The six countries in group A have the most effective systems: their administrative records are coordinated by the population registers. The four countries in group B are in an intermediate position. In the three countries in group C population registers exist only at the local level and their quality is sometimes poor. Finally Ireland and the United Kingdom are in group D at the least developed end of the spectrum. If the United Kingdom were to take what I believe is a rational and realistic course and move into group A, it would not be joining a totalitarian company.

Table 1
Particular Features of Population Registration in 15 Countries¹

	Local Population Registers	A Central Population Register which Coordinates Administrative Records	Personal Reference Numbers
A. With a Full System of Population Registration			
Belgium	x	x	x
Denmark	x	x	x
Finland	x	x	x
Luxembourg	x	x	x
Norway	x	x	x
Sweden	x	x	x
B. Intermediate Group			
France	•	x	x
Netherlands	x	•	x
Portugal	•	x	x
Spain	x	x	x
C. With Local Population registers only			
F. R. of Germany	x	•	•
Greece	x	•	•
Italy	x	•	•
D. Without Population Registers			
Ireland	•	•	•
United Kingdom	•	•	•
Number of Countries with the Feature	11	8+	10

¹ For details see Redfern 1987.

The statement noted earlier (item 6) that a properly coordinated record system will lead to better statistics needs to be qualified. Better statistics are indeed the *direct* consequence; a good example is regular and reliable population statistics for small areas. But if, as an indirect consequence, irresistible pressure builds up to replace a conventional census by a wholly register-based census, there are both benefits and penalties. Against the benefits of lower costs, a smaller burden on the public and a lesser risk of sabotage has to be set the probable deterioration in the range and quality of census results on economic topics, housing etc. Thus administrative records may increasingly fail to reflect the complexities and informalities of present-day life-styles which a conventional census could attempt to record – for example more part-time employment and self-employment, more second homes and looser family and household ties. It is here that Nordic experience (Section 3) is relevant.

Statisticians are not likely to underestimate the value of better statistics. But policy and administration – political considerations – carry a bigger weight in the arguments for and against population registers. The arguments need therefore to be debated by policy-makers,

politicians and the public. In the United Kingdom a debate ought to take place on the wisdom – indeed the feasibility – of constructing the single-purpose CC population register deliberately disconnected from other registers, rather than a multi-purpose population register with all the benefits that that could bring.

But I believe it right to bring the subject before statisticians for three reasons. First statisticians understand both the technical problems and the wider issues, and so can give a lead. Thus, in the United Kingdom both the earlier initiatives on population registers were taken in a statistical-cum-registration context (Section 5). Second, statistical agencies may be given responsibility for the key coordinating mechanisms, in particular the central population register, as INSEE has in France and SSB in Norway. Third, statisticians would benefit from more reliable data.

I hope therefore that statisticians will make their views known. Registers are very much a live issue, not least in such “under-developed” countries as the United Kingdom and Australia. Statisticians working in government service should reflect on the comment on professional ethics offered to the US Bureau of the Census; the words were written in a different context by the 1984 Panel on Decennial Census Methodology (Citro and Cohen 1985) but are very relevant here:

“We recognise that the temper of the times is not conducive to the initiation of new programs, but we believe that statisticians have the responsibility to describe the facts and recommend the actions they believe are sensible.”

ACKNOWLEDGEMENTS

For the information used in preparing this paper I am grateful to the statistical offices of Australia, Finland and Norway and, of course, of the countries which contributed to my report to the EEC. Responsibility for errors and shortcomings is mine.

REFERENCES

- AUSTRALIAN HOUSE of REPRESENTATIVES (1986). The Honorable Neal Blewett MP in the Second Reading Debate on the Australia Card Bill, 1986.
- BOREHAM, J. (1985). Quoted in How Whitehall plays the Numbers Game, *The Times*, London, 30 July 1985.
- CHARTERED INSTITUTE of PUBLIC FINANCE and ACCOUNTANCY (1987). Preparation of a specification of user requirements for the system of community charge in Scotland. CIPFA Services, London.
- CITRO, C.F., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.
- COUNCIL OF EUROPE (1981). The convention for the protection of individuals with regard to automatic processing of personal data.
- GOVERNMENT STATISTICAL SERVICE (1984). The Government Statistical Service code of practice on the handling of data obtained from statistical inquiries. Cmnd 9270, Her Majesty's Stationery Office.
- HEINONEN, R., and LAIHONEN, A. (1987). Some new solutions and methods for census data production: Finnish experiences from the 1985 census. Paper presented at the ECE/CES Seminar on Computer-Related Aspects of Population and Housing Censuses, Belgrade.

- HELDAL, J., SWENSEN, A.R., and THOMSEN, I. (1987). Census Statistics through combined use of surveys and registers? *Statistical Journal of the United Nations Economic Commission for Europe*, 5, 43-51.
- HENNESSY, P. (1987). Why journalists should breach the wall of political secrecy. *The Independent*, London, 1 April 1987.
- HER MAJESTY'S GOVERNMENT (1986). Paying for local government. Cmnd 9714, Her Majesty's Stationery Office.
- HOUSE OF LORDS (1969). The Lord Chancellor, Lord Gardiner, in Hansard, 3 December 1969.
- JENSEN, P. (1983). Towards a register-based statistical system — some Danish experience. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.
- JOHANSSON, S. (1987). Statistics based on administrative records as a substitute or a valid alternative to a population census. Paper presented at the meeting of the International Statistical Institute, Tokyo.
- LAIHONEN, A., and MYRSKYLÄ, P. (1987). Use of registers and administrative records in population censuses in Finland. Paper presented at the European Population Conference, Jyväskylä, Finland.
- MALLET, B. (1917). The organization of registration in its bearing on vital statistics. *Journal of the Royal Statistical Society*, Part 1, 80, 1-24.
- MALLET, B. (1929). Reform of vital statistics: outline of a system of national registration. *Eugenics Review*, 21, 87-94.
- PENRICE, G., REDFERN, P., EVANS, D., WHITEHEAD, F.E., BISHOP, H.E., and RUDOE, W. (1968). Discussion of the papers on social and medical statistics. *Journal of the Royal Statistical Society*, Ser. A, 131, 26-33.
- PLOWDEN, W. (1987). The battles of ideology that ill serve the public. In *The Independent*, London, 24 June 1987.
- RATING AND VALUATION ASSOCIATION (1987). Community charge, poll tax: the facts. Rating and Valuation Association, London.
- REDFERN, P. (1987). A study on the future of the census of population: alternative approaches. Eurostat Theme 3 Series C, Office for Official Publications of the European Communities, Luxembourg.

Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage

WILLIAM E. WINKLER¹

ABSTRACT

Let $A \times B$ be the product space of two sets A and B which is divided into **matches** (pairs representing the same entity) and **nonmatches** (pairs representing different entities). Linkage rules are those that divide $A \times B$ into **links** (designated matches), **possible links** (pairs for which we delay a decision), and **nonlinks** (designated nonmatches). Under fixed bounds on the error rates, Fellegi and Sunter (1969) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain probabilities that are used in a crucial likelihood ratio. In applying the record linkage model, an independence assumption is often made that allows estimation of the probabilities. If the assumption is not met, then a record linkage procedure using estimates computed under the assumption may not be optimal. This paper contains an examination of methods for adjusting linkage rules when the independence assumption is not valid. The presentation takes the form of an empirical analysis of lists of businesses for which the truth of matches is known. The number of possible links obtained using standard and adjusted computational procedures may be dependent on different samples. Bootstrap methods (Efron 1987) are used to examine the variation due to different samples.

KEY WORDS: Decision rule; Error rate; Steepest ascent; Bootstrap; Capture-recapture.

1. INTRODUCTION

This paper presents an analysis of decision rules obtained by applying the Fellegi-Sunter model of record linkage to lists of businesses. The analysis compares a rule obtained under an independence assumption that is typically assumed in practice with rules that include methods for adjusting for the failure of the independence assumption.

Given two lists, we wish to use identifying information to delineate those record pairs that represent the same entities (**matches**) and those that are different (**nonmatches**). Thus, we desire to define a linkage rule that allows us to divide the cross-product space of pairs into **links** (designated matches), **possible links** (pairs for which a decision is delayed), and **nonlinks** (designated nonmatches).

Under fixed bounds on the numbers of erroneous matches and nonmatches, Fellegi and Sunter (1969, Theorem) provide a procedure that, in theory, minimizes the number of possible links. The optimality is dependent on knowledge of certain probabilities that are used in a crucial likelihood ratio.

In typical applications, an independence assumption is made that allows estimation of the probabilities used in the likelihood ratio. The probabilities are called **matching parameters**. If the independence assumption is not valid (Winkler 1985c; Kelley 1986) then linkage rules based on the estimated probabilities may not be optimal.

¹ William E. Winkler, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233, USA.

Given fixed bounds on error rates, **better** linkage rules will be those that reduce the set of possible links. If a rule is based on matching parameters that are estimated under an invalid independence assumption, then it may be possible to develop adjustment procedures to determine better rules. To test whether one rule is statistically better than another, we use Efron's bootstrap (1987; also Hall 1988).

The remainder of the paper presents background, methods, and results from applying several record linkage rules to lists of businesses. The application involves pairs of lists for which the truth and falsehood of linkages are known.

The second section of this paper is divided into four subsections. The first contains a description of the data base and the specific subfields that are compared. The second subsection contains a summary of the Fellegi-Sunter model. The third subsection highlights common assumptions made and computational procedures used. It also contains details of computational procedures that are specific to the application of this paper.

The fourth subsection describes the evaluation procedures. The basic evaluation technique involves comparing sizes of the regions of possible links when different types of linkage rules are applied under fixed error bounds. The sizes of the regions of possible links are statistics that may be dependent on the samples used in calibrating the linkage rules. Efron's bootstrap (1987, 1982, 1979; also Hall 1988) is used to evaluate their distributions.

Results are presented in the third section. This is followed in the fourth section by discussion of the robustness of weight adjustment procedures, the type of conditioning represented by the adjusted weights, additional types of comparisons, and the use of extra blocking criteria. Finally, the paper concludes with a summary.

2. DATA BASE, LINKAGE MODEL, COMPUTATIONAL AND EVALUATION PROCEDURES

2.1 Data Base

The description of the data base is divided into two components. The first component is a description of the overall properties. The second contains a listing of the specific subfield comparisons that are made.

2.1.1 Overall Description

The data base of 57,900 records contains 54,850 records that are identified as individual companies and 3,050 duplicates. A pair of records that consists of a company and its corresponding duplicate is a match; all others are nonmatches.

The data base was constructed from 11 Energy Information Administration (EIA) and 47 State and industry lists containing 176,000 records. Duplicates were identified via elementary techniques, through call-backs (phone numbers are sometimes present) and through surveying.

The decision rules that are developed are only applied to those pairs that generally represent hard-to-identify duplicates. Easy-to-identify duplicates are those pairs having substantial portions of their name and addresses agreeing on a character-by-character basis.

An example of a hard-to-identify duplicate might be:

NAME	STREET	CITY	STATE	ZIP
Zabrinsky Fuel	16 W Sycamore St	Dayton	OH	53315
Zabrinsky Cmpny	167 Sycamere St	Springfield	OH	53315.

We observe that both ‘Zabrinsky’ and ‘Sycamore’ are spelled wrong in the second record, that ‘Cmpny’ is a nonstandard abbreviation, and that Springfield OH, a suburb of Dayton, has Postal ZIP code 53315.

2.1.2 Specific Subfields Compared

There are four sets of specific subfields that are compared in each pair of records. First are those that can be obtained through easy substring comparisons. For instance, we could compare character positions 1–4 of the NAME field from one record with the corresponding same character positions of the NAME field in another record.

In Table 1 WL-NAME is obtained by sorting the NAME field by words of decreasing length with ties broken by an alpha sort. Corresponding subfields are then compared on a character-by-character basis.

The second set is the four comparisons of the first and second largest words in the NAME field. Ties are again broken by an alpha sort.

The last two sets are of subsets of the STREET and NAME fields that are designated by highly sophisticated software. ZIPSTAN software from the Census Bureau (U.S. Dept. of Commerce 1978b) is used to obtain corresponding subfields of the STREET field. The subfields are: House No., Prefixes 1 and 2, Street Name, Suffixes 1 and 2, and Unit. Prefixes are directions such as East and North. Suffixes are words such as Street and Road. Unit designates identifiers such as apartment or suite number.

The NSKGEN5 module from software used in the Canadian Business Register (Statistics Canada 1984, 1982) is used to obtain corresponding subfields of the NAME field. NSKGEN5 creates three groups of words. The first group consists of three abbreviations with the first corresponding to surname if present. The second group contains two words with the first corresponding to surname. The third group is a single word obtained by concatenating and abbreviating individual words in the NAME field. Details are given in Winkler (1987) or in Statistics Canada (1984, 1982).

2.2 Fellegi-Sunter Model

The Fellegi-Sunter Model uses a decision-theoretic approach establishing the validity of principles first used in practice by Newcombe (Newcombe *et al.* 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The description closely follows Fellegi and Sunter (1969, pp. 1184-1187).

Table 1
Corresponding Subfields Compared on a
Character-by-Character Basis

Field	1-4, 5-10, 11-20, 21-30
NAME	1-4, 5-10, 11-20, 21-30
STREET	1-6, 7-15, 16-30
ZIP	1-3, 4-5
CITY	1-5, 6-10, 11-15
STATE	1-2
TELEPHONE	1-3, 4-6, 7-10
WL-NAME	1-4, 5-10, 11-20, 21-30

There are two populations A and B whose elements will be denoted by a and b . We assume that some elements are common to A and B . Consequently the set of ordered pairs

$$A \times B = \{(a,b): a \in A, b \in B\}$$

is the union of two disjoint sets of **matches**

$$M = \{(a,b): a = b, a \in A, b \in B\}$$

and **nonmatches**

$$U = \{(a,b): a \neq b, a \in A, b \in B\}.$$

The records corresponding to members of A and B are denoted by $\alpha(a)$ and $\beta(b)$, respectively. The **comparison vector** γ associated with the records is defined by:

$$\gamma[\alpha(a), \beta(b)] \equiv \{\gamma^1[\alpha(a), \beta(b)], \gamma^2[\alpha(a), \beta(b)], \dots, \gamma^K[\alpha(a), \beta(b)]\}.$$

Each of the $\gamma^i, i = 1, \dots, K$, represents a specific comparison. For instance, γ^1 could represent agreement/disagreement on sex. γ^2 could represent the comparison that two surnames agree and take a specific value or that they disagree.

Where confusion does not arise, the function γ on $A \times B$ will be denoted by $\gamma(\alpha, \beta), \gamma(a, b)$, or γ . The set of all possible realizations of γ is denoted by Γ .

The conditional probability of $\gamma(a, b)$ if $(a, b) \in M$ is given by

$$\begin{aligned} m(\gamma) &\equiv P\{\gamma[\alpha(a)\beta(b)] | (a,b) \in M\} \\ &= \sum_{(a,b) \in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a,b) | M]. \end{aligned}$$

Similarly we denote the conditional probability of γ if $(a,b) \in U$ by $u(\gamma)$.

We observe a vector of information $\gamma(a, b)$ associated with pair (a, b) and wish to designate a pair as a link (denote the decision by A_1), a possible link (decision A_2), or a nonlink (decision A_3). A **linkage rule** L is defined a mapping from Γ , the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

$$d(\gamma) = \{P(A_1 | \gamma), P(A_2 | \gamma), P(A_3 | \gamma)\}; \gamma \in \Gamma$$

and

$$\sum_{i=1}^3 P(A_i | \gamma) = 1.$$

There are two types of error associated with a linkage rule. A **Type I error** occurs if an unmatched comparison is erroneously linked. It has probability

$$P(A_1 | U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1 | \gamma)$$

A **Type II error** occurs if a matched comparison is erroneously not linked. It has probability

$$P(A_3|U) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3|\gamma).$$

Fellegi and Sunter (1969) define a linkage rule L_0 , with associated decisions A_1 , A_2 , and A_3 , that is optimal in the following sense:

THEOREM (Fellegi-Sunter 1969). Let L' be a linkage rule with associated decisions A'_1 , A'_2 , and A'_3 such that it has the same error probabilities $P(A'_3|M) = P(A_3|M)$ and $P(A'_1|U) = P(A_1|U)$ as L_0 . Then L_0 is optimal in that $P(A_2|U) \leq P(A'_2|U)$ and $P(A_2|M) \leq P(A'_2|M)$.

In other words, if L' is any competitor of L_0 having the same Type I and Type II error rates (which are both conditional probabilities), then the conditional probabilities (either on set U or M) of not making a decision under rule L' are always greater than under L_0 . L_0 is described in subsection 2.3.1.

The Fellegi-Sunter linkage rule is actually optimal with respect to any set Q of ordered pairs in $A \times B$ if we define error probabilities P_Q and a linkage rule L_Q conditional on Q . Thus, it may be possible to define subsets of $A \times B$ on which we make use of differing amounts and types of available information.

For instance, if we have a set of pairs in which telephone number is present, we might use telephone number and a few characters from the name to designate links. With other pairs, we may additionally have to utilize information from the street address and the city name.

Sets of ordered pairs Q on which the Fellegi-Sunter linkage rule is applied are often obtained by **blocking criteria**. Blocking criteria are sort keys that are used to reduce the number of pairs that are considered. Rather than consider all pairs in $A \times B$, we might only consider pairs that agree on the first three digits of the ZIP code or on a suitable abbreviation of surname.

2.3 Computational Procedures

This section is divided into five parts. The first part contains a description of the general linkage rule of the Fellegi-Sunter Model. The second contains a description of the simplified computational procedures when a conditional independence assumption is made.

Background on the validity of the conditional independence assumption is presented in the third part. The fourth describes two general methods of adapting computational procedures. The fifth provides a description of the specific computational procedures of this paper.

2.3.1 General Form of Linkage Rule

To provide a background for understanding why specific computational procedures are used, we consider the following likelihood ratio

$$R \equiv R[\gamma(a,b)] = m(\gamma)/u(\gamma). \quad (2.1)$$

We observe that, if γ represents a comparison of K fields, then there are at least 2^K probabilities of form $m(\gamma)$. If γ represents agreements of K fields, we would expect this to occur more often for matches M than for nonmatches U . The ratio R would then be large. Alternatively, if γ consists of disagreements, the ratio R would be small.

If the numerator is positive and the denominator is zero in (2.1), we assign an arbitrary very large number to the ratio. The Fellegi-Sunter linkage rule takes the form:

$$\begin{aligned} \text{If } R > \text{UPPER, then denote } (a,b) \text{ as a link.} \\ \text{If } \text{LOWER} \leq R \leq \text{UPPER, then denote } (a,b) \text{ as a possible link.} \\ \text{If } R < \text{LOWER, then denote } (a,b) \text{ as a nonlink.} \end{aligned} \quad (2.2)$$

The cutoffs LOWER and UPPER are determined by the desired error rate bounds.

2.3.2 Simplification Under Conditional Independence Assumption

In practice, computation is simplified two ways. The first is by the conditional independence assumption of Fellegi and Sunter (1969):

For each $\gamma \in \Gamma$

$$\begin{aligned} m(\gamma) &= m_1(\gamma^1) \cdot m_2(\gamma^2) \dots m_K(\gamma^K) \text{ and} \\ u(\gamma) &= u_1(\gamma^1) \cdot u_2(\gamma^2) \dots u_K(\gamma^K) \end{aligned}$$

where for $i = 1, 2, \dots, K$

$$\begin{aligned} m_i(\gamma^i) &= P(\gamma^i | (a,b) \in M) \text{ and} \\ u_i(\gamma^i) &= P(\gamma^i | (a,b) \in U). \end{aligned}$$

This assumption basically is that agreement on one characteristic such as surname does not depend on agreement of other characteristics such as house number or age.

The second is to use a computationally convenient function of the ratio in (2.1). Log_2 is used. We then have

$$\begin{aligned} W &\equiv W(\gamma) = \text{Log}_2[m(\gamma)/u(\gamma)] \\ &= W^1 + W^2 + \dots + W^K, \end{aligned} \quad (2.3)$$

where $W^i \equiv \text{Log}_2[m_i(\gamma^i)/u_i(\gamma^i)]$ for $i = 1, 2, \dots, K$. We call W the **total comparison weight** associated with a pair and W^i , $i = 1, 2, \dots, K$, the **individual comparison weights**.

For the remainder of the paper we will assume that each component γ^i , $i = 1, 2, \dots, K$, in γ represents a two-state comparison (e.g., agree/disagree). For convenience, we denote agreement in the i th component by γ_o^i , $i = 1, 2, \dots, K$. Under the conditional independence assumption, for each $i = 1, 2, \dots, K$, we need to estimate probabilities of the forms

$$P(\gamma = \gamma_o^i | M) \text{ and } P(\gamma = \gamma_o^i | U). \quad (2.4)$$

Using a set of pairs for which the truth and falsehood of matches are known, for each agreement γ_o^i , $i = 1, 2, \dots, K$, we divide the set into the four subsets determined by the agree/disagree and match/nonmatch statuses in (2.4) to perform the estimation.

If no conditional independence assumption is made, we need to estimate $2 \cdot (2^K - 1)$ probabilities of form (2.1) and divide the set of pairs for which truth and falsehood are known to $2 \cdot (2^K - 1)$ subsets. Even with a small number of comparisons (say, 6 or less), we may not be able to obtain sufficiently large samples to allow accurate estimation of the probabilities.

2.3.3 Validity of Conditional Independence Assumption

Winkler (1985c) has shown that the independence assumption is not valid for simple comparisons of portions of the name and street address fields for list of businesses. Using similar portions of the name and street fields, Kelley (1986) has shown that the independence assumption is not valid for files of individuals. Furthermore, Kelley and Winkler have each shown that matching efficacy is sensitive to the set of pairs over which probabilities of the form (2.4) are computed.

Fellegi and Sunter indicate that, if the conditional independence assumption is not valid, then estimates of weights that are obtained via formula (2.3) will lose their strict probabilistic interpretation. By this, they mean that the linkage rule of their theorem may not actually minimize the number of possible links. They indicate that they believe their procedure to be robust to departures from the independence assumption.

Under the independence assumption, probabilities are computed as products of probabilities of the form (2.4). If we have a set of pairs for which truth and falsehood of matches are known, then we can adjust probabilities of form (2.4) for departures from the independence assumption. If the total weights obtained by adjustment yield substantially smaller sets of potential links under fixed bounds on error rates, then the Fellegi-Sunter procedure may not be robust to departures from independence.

2.3.4 General Adjustments

There are two general adjustments to the basic methods of computing individual comparison weights. The first consists of dividing the subset of pairs in $A \times B$ over which individual comparison weights are computed into several subsets. The linkage rule is obtained by restricting the basic Fellegi-Sunter rule to correspond to the different subsets on which weights are computed. Individual comparison weights may vary significantly in different subsets.

The second adjustment consists of modifying individual comparison weights. Under the independence assumption, we consider the equation

$$\begin{aligned} W &\equiv \text{Log}_2(P(\gamma \in B_1 \cap B_2 \cap \dots \cap B_K | M) / P(\gamma \in B_1 \cap B_2 \cap \dots \cap B_K | U)) \\ &= W^1 + W^2 + \dots + W^K, \end{aligned}$$

where, for $i = 1, 2$, and K , $W^i \equiv \text{Log}_2(P(\gamma \in B_i | M) / P(\gamma \in B_i | U))$ and B^i is the set $\{\gamma^i = \gamma_0^i\}$ or its complement. We wish to find computationally tractable methods of adjusting the W^i , $i = 1, 2, \dots, K$, so that their sum yields better linkage rules.

If there is a sample for which the truth and falsehood of matches are known, then we can estimate individual comparison weights (Tepping 1968) and the adjustments.

The simplest adjustment procedure involves a steepest ascent approach (e.g., Cochran and Cox 1957). To begin, we use the known truth and falsehood of matches within a sample to estimate probabilities of the form (2.4). The probabilities are then used in computing individual comparison weights that are added to obtain an estimate of total weight (2.3). For each pair of fixed bounds on Type I and Type II errors, the cutoffs UPPER and LOWER of (2.2) can be determined. The number of potential links for rules of the form (2.2) follows immediately.

Next, we chose an individual comparison weight, change it by a fixed amount (say ± 1), recompute the total weight (2.3) using the new individual weight, and find new cutoffs UPPER and LOWER and a new region of potential links.

If under fixed bounds of errors, the size of the region of possible links decreases, then we continue adjusting the individual comparison weight (either up or down) until the region ceases its decrease in size. We continue by varying other individual weights in a similar manner.

If the size of the region of possible links decreases substantially, then we know the conditional independence assumption is not valid for the set of comparisons. If the conditional independence assumption were valid, then the estimated weights would accurately represent the true weights. The regions of possible links would be minimal by the theorem of Fellegi and Sunter.

A linkage rule that is based on adjusted individual comparison weights depends on the sample used in the steepest ascent procedure.

2.3.5 Specific Methods

To describe the specific methods of computing weights and obtaining corresponding linkage rules used in this paper, we need some additional background.

The only pairs considered are those that agree on at least one of the blocking criteria in Table 2.

We subdivide the set of pairs obtained via the four sets of blocking criteria into the five classes given in Table 3.

Table 2
Blocking Criteria

#	Characters Used
1.	3 digits ZIP, 4 characters NAME
2.	5 digits ZIP, 6 characters STREET
3.	10 digits TELEPHONE
4.*	Word length sort NAME field, then use 1.

* This criterion also has a deletion stage which prevents matching on commonly occurring words such as 'OIL', 'FUEL', 'CORP', and 'DISTRIBUTOR.'

Table 3
Sets of Pairs Determined by Blocking Criteria

Class	# pairs	Determining Blocking Criteria
1	1021	Agreeing on criterion 1 and no other or simultaneously agreeing on criteria 1 and 4 and no others.
2	624	Agreeing on criterion 2 and no other or simultaneously agreeing on criteria 2 and 3 and no others.
3	256	Agreeing on criterion 3 only.
4	344	Agreeing on criterion 4 only.
5	2240	Agreeing on at least one criterion but not in classes 1-4.

Class 5 contains pairs that generally agree on two or more blocking criteria. Classes 1-5 contain 2991 matches and 1494 nonmatches and miss 59 known matches. The determination of sets of blocking criteria and classes is treated in detail in Winkler (1985b, 1987).

We classify linkage rules by the different ways in which the individual comparison weights are computed and how resultant linkage rules are defined.

The first type, AA, of weight computation is an overall aggregate in all pairs. The second, A, is an overall aggregate in classes 1-4. The third, U, yields separate weight computations in classes 1-4. The fourth, C, uses steepest ascent to adjust the individual weight computation of Type U.

Each successive type of linkage rule involves increasingly more complex weight computations. Matches outside classes 1-5 are not considered in the results section because their number is constant for each of the four linkage rules.

2.4 Evaluation Procedures

The basic evaluation technique involves comparing sizes of the region of possible links when the different types of linkage rules are applied under fixed error bounds.

Efron's bootstrap (1987, 1982, 1979) is used to estimate confidence intervals for statistics such as the number of possible links. As these statistics are obtained under complicated rules, it seems unlikely that closed-form estimates can be determined.

If there are sets of pairs for which the truth and falsehood of matches are known, then we can use Efron's bootstrap to estimate the variation of parameters in the following fashion:

- 1. Draw calibration samples of equal size with replacement.
- 2. Estimate individual comparison weights of the form (2.4) using the known truth and falsehood in the sample and use them to estimate total weight via (2.3).
- 3. Compute cutoffs LOWER and UPPER using each sample (in our application we allow at most 2 percent of the links to be nonmatches and 3 percent of the nonlinks to be matches).
- 4. Using individual comparison weights from step 2, compute a total comparison weight for each pair in the entire selected set of pairs. Use cutoffs from step 2 to classify pairs as links, possible links, and nonlinks.
- 5. Using estimates from individual samples, determine the means and variances of the cutoff weights, of the misclassification rates, and of the number of possible links.

The bounds (2 and 3 percent, step 3) are used to try to assure that the corresponding classification error rates in the entire data base are less than 5 percent.

Table 4
Linkage Rules by Type of Weight Computation and
Sets of Pairs to Which Applied

Type	Individual Weight Computation	Linkage Rule
AA	Uniformly over all pairs in Classes 1-5	Over all pairs
A	Uniformly over all pairs in Classes 1-4	Designate pairs in Class 5 Links, Apply Fellegi-Sunter Rule to remaining pairs in Classes 1-4
U	Uniformly in each Class 1-4	Designate pairs in Class 5 Links, Apply Fellegi-Sunter Rule individually in Classes 1-4
C	Uniformly in each Class 1-4	Same as U except modify weights using steepest ascent procedure

Computations and adjustments must be performed consistently across calibration samples. Identical adjustment procedures must be used in obtaining individual adjusted weights, total weights, and cutoffs. If an individual weight is adjusted upward (step 2) by amount x or percentage y with one sample, then the same adjustment must be used with other samples.

As the underlying distributions may not be normal or may be biased and skewed, we can use new techniques of Efron (1982, 1987; also Hall 1988) to determine confidence intervals. Hall (1988) has shown the theoretical validity of the nonparametric bootstrap that includes an acceleration-constant type adjustment for skewness of a distribution.

3. RESULTS

The results in this section comprise three parts. The first part is an overall comparison from using the four different weighting methods described in section 2.3.5. The second part contains more details about the best two methods from the first part. The third part contains results from the bootstrap evaluation.

3.1 Overall Comparison

We place fixed upper bounds of 5 percent on the number of matches misclassified as nonmatches and 2 percent on the number of nonmatches misclassified as matches. As we are using discrete data, actual error rates will generally not equal their upper bounds (Table 5, columns 2 and 3).

We see that, as the complexity of the application of the weighting methodology increases, the number of possible links (size of manual review region) decreases dramatically from 1512 to 97. This indicates that the increasing complexity of the weight computations yields increasingly better decision rules.

We see that the last two methods, which both involve computing individual comparison weights separately in classes 1–4, yield the smallest sets of possible links (695 and 97, respectively).

3.2 Best Methods

We consider the best two methods, linkage rules using weights of Type U and of Type C, in greater detail. Results from applying weights of Type U and Type C are presented in Tables 6 and 7, respectively. In determining cutoff weights by class, we place rough upper bounds of 5 percent misclassified nonmatches and 2 percent misclassified matches in each class. The overall upper bound is maintained.

Comparing columns 4 and 5 across tables 6 and 7, we find that the corresponding numbers of misclassified matches and nonmatches are approximately the same. This is consistent with the bounding method. In every class, the linkage rule using Type C weights yields less possible links than the rule using Type U weights.

The numbers of records classified as possible links are less in classes 1 and 4 (83 versus 55 and 44 versus 0, respectively) and dramatically less in classes 2 and 3 (409 versus 0 and 159 versus 42, respectively).

One hundred percent of the pairs in classes 2 and 4 are classified by the procedure that uses Type C weights.

Two variations distinguish the linkage rule based on type C weights from the rule based on type U weights. First, we vary agreement weights associated with the four subfields of the NAME after words have been sorted by decreasing length (Table 8). The only substantial variations (greater than 2.5 on the \log_2 scale) occur in Class 2.

Table 5
Error Rates and Number of Possible Links
from Applying Different Weighting Methods

Weight Type	Proportion Misclassified as		Total Classified		Possible Links
	Non-Match	Match	Non-Match	Match	
AA	.047	.020	964	2009	1512
A	.041	.015	952	2481	1052
U	.050	.020	1083	2707	695
C	.033	.019	1441	2947	97

Table 6
Results from Using a Linkage Rule Based on Type U
Weights for Delineating Matches and Nonmatches
(5 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassified as		Total Classified as		Total Not Classified	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	0.5	6.5	39	14	674	264	83	1021
2	-4.5	3.5	2	4	100	115	409	624
3	-4.5	6.5	2	1	55	42	159	256
4	2.5	11.5	11	2	254	46	44	344
Totals			54	21	1083	467	695	2245

Table 7
Results from Using a Linkage Rule Based on Type C
Weights for Delineating Matches and Nonmatches
(3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassified as		Total Classified as		Total Not Classified	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	55	1021
2	2.5	2.5	5	3	379	245	0	624
3	-0.5	4.5	5	6	104	110	42	256
4	8.5	8.5	9	4	266	78	0	344
Totals			47	21	1441	707	97	2245

Table 8
Steepest Ascent Adjustment to Agreement Weights
for Subfields Obtained by Wordlength Sort¹

Class	Subfield			
	1	2	3	4
1	.	.	—	+
2	++	++	+	+
3	+	+	—	++
4	.	+	—	+

¹ '.' means deviation less than 1.0, '+', '—' mean deviation greater than 1.0 and less than 2.5, and '++' means deviation greater than 2.5.

The second is that the agreement weight is only utilized if four corresponding subfields, the three subfields of CITY and the one STATE, agree. The variation, in effect, typically increases the **relative** distinguishing power of agreements/disagreements in subfields other than the CITY field.

The largest reduction (from 409 to 0) in the number of possible links takes place in Class 2. A slightly higher proportion ($.95 \approx 359/379$) of nonlinks have an agreeing CITY field than links ($.91 \approx 223/245$).

The following is an example of a match that is not designated as a link using the rule based on Type U weights but is using the rule based on Type C weights.

NAME	STREET	CITY	STATE	ZIP
Roberts Heat Oils	167 Sycamore St	Dayton	OH	53315
Maxwell S Robert Heat Oil	167 Sycamore St	Dayton	OH	53315.

The first six digits of the telephone number also agreed.

The following is an example of an erroneous match using Type C weights.

NAME	STREET	CITY	STATE	ZIP
Molar Petro	167 Sycamore St	Dayton	OH	53315
Petrochem	167 Sycamore St	Dayton	OH	53315.

These two companies do business from the same location and also have identical phone numbers.

The following is an example of an erroneous nonmatch using Type C weights.

NAME	STREET	CITY	STATE	ZIP
Johns Geo M	167 Sycamore St	Springfield	OH	53315
Geo M Johns Jobber	167 Sycamore	Spring Field	OH	53315.

Insertion or deletion of blanks in corresponding fields typically causes record pairs to be designated as a nonmatch.

Table 9
Bootstrap 90 Percent Confidence Intervals for Counts of Possible Links
500 Replications

Weight Type	Class	Ordinary Interval	BC Interval	BC _a Interval
C	1	(42,117)	(37,108)	(37,108)
C	2	(0, 0)	(7, 7)	(7, 7)
C	3	(31,154)	(34,156)	(34,156)
C	4	(0, 36)	(0, 39)	(0, 39)
U	1	(122,192)	(128,196)	(128,196)
U	2	(383,501)	(383,501)	(383,501)
U	3	(149,201)	(142,197)	(142,197)
U	4	(35, 82)	(33, 81)	(33, 81)

3.3 Bootstrap Variation

The results of this section involve increasingly more sophisticated methods of computing bootstrap confidence intervals (Table 9). For each class, 500 replications are used in computing 90 percent confidence intervals for estimates of the number of records designated as possible links. The two error bounds are fixed at 5 percent.

The first interval is the ordinary bootstrap interval that is partially based on normal theory (Efron 1979). The second interval, denoted by BC, is an interval in which a bias adjustment has been made (Efron 1979, 1982). The third interval, denoted by BC_a, is obtained using acceleration-constant type adjustments for bias and skewness (Efron 1987; also Hall 1988).

Examination of Table 9 yields that each of the intervals in respective classes are approximately the same length. If the method of adjusting to achieve weights of Type C were highly sensitive to the individual samples taken for calibration, we would expect the confidence intervals associated with Type C weights to be larger than those associated with Type U weights.

The fact that the intervals are large for either type of weight indicates the results are quite dependent on the calibrating samples. The fact that the ordinary confidence intervals are roughly the same as the BC and BC_a indicates that the respective distributions are neither biased nor skewed.

The number of possible links in intervals based on Type C weights is almost always less than the corresponding intervals based on Type U weights. Only the intervals associated with classes 3 and 4 show slight overlap. Thus, it is reasonable to accept the hypothesis that the linkage rule based on Type C weights consistently outperforms the linkage rule based on Type U weights.

4. DISCUSSION

This section is composed of four parts. The first contains a discussion of the robustness of the steepest ascent adjustments. The second subsection describes the implicit type of conditioning imposed by the steepest ascent adjustments. The third part considers the usefulness of making comparisons that are partially dependent on other comparisons. The fourth subsection describes methods for determining sets of blocking criteria.

4.1 Robustness of Steepest Ascent Adjustment

The sizes of regions of possible links are somewhat sensitive to the set of weights that are varied during the steepest ascent procedure. In two cases (one of which was presented in this paper), the numbers of possible links were approximately 100; in two others, 200. All four of the steepest ascent variations yielded improvements over the 700 possible links obtained by the best non-steepest ascent procedure.

The individual weights that were modified varied significantly over the four cases. In no case were more than eight of the 30 weights varied.

It is reasonable to hypothesize that the steepest ascent weighting procedure will yield improvements when deviations from conditional independence are substantial. No bootstrap-based significance tests were used to check the hypothesis for three of the four cases.

Obtaining small samples that allow adjustments such as performed in this paper should be straightforward. Sample sizes of 100 in each class may be sufficient. The sample sizes used for the bootstrap results of section 3.3 were approximately 100 in each class. Comparable bootstrap results using samples of 30 and 50 in each class were not sufficient to show that adjustments yielded quantifiable improvements. Sample sizes of 200 yielded bootstrap confidence intervals that were almost the same as those based on samples of sizes 100.

Many record linkage systems (*e.g.*, U.S. Dept. Agriculture 1979; U.S. Dept. of Commerce 1978a; Statistics Canada 1984) allow modification of matching parameters based on information from samples. Reestimation of parameters using sample information is a powerful feature of the Generalized Iterative Record Linkage System of Statistics Canada (1983). The parameter-reestimation in these systems generally involves direct reestimation of the marginal probabilities $m_i(\gamma^i)$ and $u_i(\gamma^i)$. It does not involve adjustments of weights such as given in this paper.

4.2 Type of Conditioning Represented by Modified Weights

To prepare for the discussion in this section, we need two sets of facts. The first set involves the conditional discriminating power of components of γ . Let σ be a vector with components $\sigma^1, \sigma^2, \dots, \sigma^K$ that consists of a reordering of the components $\gamma^1, \gamma^2, \dots, \gamma^K$ of γ . Then

$$\begin{aligned} P(\gamma|M) &= P(\sigma|M) = \\ P(\sigma^1 = \sigma_0^1, \sigma^2 = \sigma_0^2, \dots, \sigma^K = \sigma_0^K | M) &= \end{aligned} \quad (4.1)$$

$$P(\sigma^1 = \sigma_0^1 | M) \cdot P(\sigma^2 = \sigma_0^2 | \sigma^1, M) \dots P(\sigma^K = \sigma_0^K | \sigma^1, \sigma^2, \dots, \sigma^{K-1}, M).$$

The component σ^1 might refer to first name, σ^2 to house number, σ^3 to age, and so on.

For each σ we can call $P(\sigma^i = \sigma_0^i | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, M)$ the **successive conditional incremental discriminating component** of σ^i in M , $i = 1, 2, \dots, K$. These incremental probability components are dependent on the reordering $\sigma^1, \sigma^2, \dots, \sigma^K$. Each component on the right hand side of (4.1) is independent of the others. In a similar manner, we can consider incremental components in U .

The basic purpose of a reordering is to consider one specific pattern of conditional probabilities for $\gamma \in \Gamma$. For the single reordering we let $\sigma = \sigma(\gamma)$ vary in $\sigma(\Gamma)$ as $\gamma \in \Gamma$. Then for all $\sigma \in \sigma(\Gamma)$,

$$\begin{aligned}
 W &\equiv W(\gamma) = \text{Log}_2[m(\gamma)/u(\gamma)] \\
 &= A^1 + A^2 + \dots + A^k,
 \end{aligned}
 \tag{4.2}$$

where $A^i \equiv \text{Log}_2[P(\sigma^i = \sigma_0^i | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, M) / P(\sigma^i = \sigma_0^i | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, U)]$ for $i = 1, 2, \dots, K$.

The second set of facts involves transformations that map the ratio R given by (2.1) to real numbers which we call **weights**. For each pair of Type I and Type II errors, we consider any transformation that places weights associated with links in the highest interval, weights associated with nonlinks in the lowest interval, and weights associated with possible links in the interval between the upper and lower intervals. Such a transformation yields rules that can be represented in forms similar to form (2.2) and are equivalent to the Fellegi-Sunter rule at the same fixed pair of error levels. If the transformation is monotone, then the new weights yield rules that are equivalent to the original Fellegi-Sunter rule for all error levels.

The steepest ascent weight adjustment procedure implicitly determines a transformation of the ratio R and a single reordering that is fixed for all $\gamma \in \Gamma$ and the same in M and U . The fact that the steepest ascent procedure adjusts weights sequentially assures that there is a single reordering. The adjusted weights $W^i \pm c_i$ are estimates that replace the W^i in (2.3) for some real constants c_i , $i = 1, 2, \dots, k$.

The fact that the adjusted weights yield smaller regions of possible links means that, at a fixed pair of error levels, the new total weights more accurately represent a transformation of the Log_2 of the ratio of the true probabilities given by the left hand of (4.1). The new total weights represent estimates that transform the right hand side of (4.2).

The adjustment procedure allows us to utilize better the incremental distinguishing power of one field given another, a second field given the first two, and so on. We note that we do not need to know the specific transformation or the specific pattern of conditioning induced by the reordering.

The adjustment procedure is similar to new bootstrap procedures (Efron 1987; Hall 1988). The validity of the bootstrap procedures is dependent on the existence of monotone transformations, bias constants, and acceleration constants that yield the exact correspondence of confidence intervals of the original distributions with confidence intervals of specified normal distributions. The transformations and constants need not be known.

4.3 Value of Dependent Comparisons

The intuitive idea of making a number of comparisons, some of which may be partially dependent on other comparisons, is that they may, when used in properly adjusted rules, yield additional distinguishing power. Newcombe and Kennedy (1962, see also Newcombe *et al.* 1983) have given examples of comparisons of portions of name fields that intuitively may be dependent on other comparisons. The additional comparisons, nevertheless, may yield better linkage rules than those rules that do not utilize the same additional comparisons.

The chief difficulty in using additional comparisons is properly utilizing their incremental distinguishing power. This paper's set of comparisons – in particular, of subfields of the name field – is not independent in the sense of equation (2.3). The primary purpose of the set is to illustrate methods for systematically obtaining better linkage rules when the conditional independence assumption is not valid.

4.4 Additional Blocking Criteria

There are two conflicting goals when a set of blocking criteria is used to reduce the number of pairs in $A \times B$ that receive further processing. The first is the need to reduce (drastically) the number of pairs that are processed and to obtain a set in which linkage rules can accurately delineate matches and nonmatches. The second is to obtain a set that contains as many matches from M as possible.

To determine whether it is feasible to look for additional sets of blocking criteria, it is first necessary to find estimates of the number of matches missed by a given set of blocking criteria. If the estimates are acceptably small, then it is not necessary to look for additional criteria.

To estimate the number of matches missed by given sets of blocking criteria, Scheuren (1983) suggested using standard capture-recapture techniques such as given in Bishop, Fienberg, and Holland (1975, Chapter 6). Winkler (1987) applied the techniques to the same empirical data and four sets of blocking criteria as in this paper.

The best fitting loglinear model for the table of counts of records captured and not captured by the four sets of blocking criteria was used in obtaining a confidence interval for the number of matches missed. Based on assumed asymptotic normality, a 95 percent confidence interval (27,160) was computed. The interval represents between 1 and 5 percent of the matches.

5. SUMMARY

The results of this paper show that the conditional independence assumption is not always valid. When the assumption is not valid, it is possible to develop adjusted linkage rules that improve on the standard linkage rule. Under fixed bounds on error rates, the improved rules reduce the size of the region of possible links.

REFERENCES

- BISHOP, Y. M. M., FIENBERG, S. E., and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.
- COCHRAN, W.G., and COX, G.M. (1957). *Experimental Designs*. New York: John Wiley and Sons.
- EFRON, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics*, 7, 1-26.
- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Methods*. Philadelphia: SIAM.
- EFRON, B. (1987). Better Bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, 82, 171-185.
- FELLEGI, I. P., and SUNTER, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.
- HALL, P. (1988). Theoretical comparison of Bootstrap confidence intervals. *Annals of Statistics*, 16, 927-953.
- KELLEY, R. P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A. P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B., and KENNEDY, J.M. (1962). Record linkage. *Communications of the Association for Computing Machinery*, 5, 563-566.
- NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual searches in a study of the health of Eldorado Uranium workers. *Computers in Biology and Medicine*, 13, 157-169.

- SCHEUREN, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-381.
- SCHEUREN, F. (1985). Methodological issues in linkage of multiple data bases, in *Record Linkage Techniques - 1985*, edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 155-167.
- STATISTICS CANADA (1982). Record Linkage Software, Systems Development Division.
- STATISTICS CANADA (1983). Generalized Iterative Record Linkage System, Systems Development Division.
- STATISTICS CANADA (1984). Record Linkage Software, EDP Planning and Support Division.
- TEPPING, B. J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association* 63, 1321-1332.
- U.S. DEPARTMENT OF AGRICULTURE (1979). List Frame Development: Procedures and Software, Statistical Reporting Service.
- U.S. BUREAU OF THE CENSUS (1978a). UNIMATCH: A Record Linkage System, Survey Research Division.
- U.S. BUREAU OF THE CENSUS (1978b). ZIPSTAN: Generalized Address Standardizer, Survey Research Division.
- WINKLER, W. E. (1985a). Preprocessing of lists and string comparison, in *Record Linkage Techniques - 1985*, edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 181-187.
- WINKLER, W. E. (1985b). Exact matching lists of businesses: Blocking, subfield identification, and Information Theory, in *Record Linkage Techniques - 1985*, edited by W. Alvey and B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 227-241.
- WINKLER, W. E. (1985c). Exact matching lists of businesses. *Proceedings of the Section on Survey Research Methods, American Statistical Association* 438-443.
- WINKLER, W. E. (1987). An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses, Energy Information Administration, Technical Report.
- WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association* to appear.

Automated Quality Assurance Processing of Administrative Record Files

JAMES R. JONAS and PAUL S. HANCZARYK¹

ABSTRACT

The Census Bureau makes extensive use of administrative records information in its various economic programs. Although the volume of records processed annually is vast, even larger numbers will be received during the census years. Census Bureau mainframe computers perform quality control (QC) tabulations on the data; however, since such a large number of QC tables are needed and resources for programming are limited and costly, a comprehensive mainframe QC system is difficult to attain. Add to this the sensitive nature of the data and the potentially very negative ramifications from erroneous data, and the need becomes quite apparent for a sophisticated quality assurance system on the microcomputer level. Such a system is being developed by the Economic Surveys Division and will be in place for the 1987 administrative records data files. The automated quality assurance system integrates micro and mainframe computer technology. Administrative records data are received weekly and processed initially through mainframe QC programs. The mainframe output is transferred to a microcomputer and formatted specifically for importation to a spreadsheet program. Systematic quality verification occurs within the spreadsheet structure, as data review, error detection, and report generation are accomplished automatically. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces processing costs on the mainframe and provides the comprehensive quality assurance component for administrative records.

KEY WORDS: Mainframe-microcomputer integration; Systematic data verification; Timeliness.

1. INTRODUCTION

The Bureau of the Census makes extensive use of administrative record information in our economic programs. The data originate from the business-related tax collection processes of the Internal Revenue Service (IRS) and, to a lesser extent, the Social Security Administration. During economic and agriculture censuses years, the volume of administrative record data received increases substantially. These data have enabled us to conduct economic and agriculture censuses on a timely and efficient basis and with a minimum of reporting burden on the business and farm communities. The success of our economic and agriculture programs depends to a great extent on the timeliness and quality of these administrative record files.

It is vital for Census Bureau operations to ensure the quality of all incoming data. As in past economic censuses, we have developed mainframe quality assurance programs for the administrative record data. However, since such a large number of these tables are needed and resources for programming are limited and costly, a comprehensive quality assurance system is difficult to attain entirely on the mainframe. Add to this the sensitive nature of these data and the potential ramifications of erroneous data, and the need for a more sophisticated quality assurance system becomes apparent. The Census Bureau has developed a comprehensive quality

¹ James R. Jonas and Paul S. Hanczaryk, Economic Surveys Division, U.S. Bureau of the Census, Washington, D.C., 20233, U.S.A.

assurance system that manages various phases of our administrative records review process. This automated system will allow us to perform more thorough quality assurance within the bounds of restrictive budgets and limited programming resources.

The automated quality assurance system integrates mainframe computer and microcomputer technology. The Census Bureau has established standards that delineate our fundamental requirements of the incoming administrative record data set. These standards are entered into a microcomputer system. After the mainframe quality assurance programs are run, the results are downloaded into the same microcomputer system. The reporting patterns of the actual administrative record data are then compared to the predetermined standards. Mechanical data verification occurs as data review, error detection, and report generation are accomplished automatically at the microcomputer level. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces the processing costs on the mainframe. Moreover, the system provides the quality assurance component needed for thorough and unerring review of administrative records. Although designed specifically for the IRS business income tax return files used in the censuses, it can and will be adapted to all incoming administrative record files after 1988.

2. OVERVIEW OF QUALITY ASSURANCE SYSTEM FROM A MANAGEMENT PERSPECTIVE

Administrative records play a major role at the Census Bureau, a role that has steadily grown in importance over time. The increasing need for more and better statistics, the need to compile those statistics with a minimum of burden on the private sector, and the need to use our available human and financial resources as efficiently as possible have all contributed to the importance of administrative records.

Over the past several years, the quality of the administrative records generally has been excellent. However, we did experience certain problems with the quality of the 1982 business income tax data from the IRS. The most detrimental problem was the inadequate quality of the principal industrial activity codes for sole proprietorships. As a result of this problem, the Census Bureau published only limited statistics for nonemployers in the 1982 Economic Censuses. If our quality assurance programs had been more sophisticated, the errors could have been identified earlier and the negative impact would have been minimized.

Heading into the 1987 Economic Censuses, it was determined that additional measures were needed to ensure the quality of administrative record data received from the IRS. An overall quality management system responsive to certain factors that have adversely affected past administrative data sets was necessary. The three major factors that have plagued us in the past are:

1. Vast amounts of administrative record data

The IRS will provide us with selected business 1987 tax return data (received in 1988) for various legal forms of businesses, including corporations, S corporations, foreign corporations, partnerships, nonprofit organizations, and sole proprietorships. In total, the Census Bureau expects over 75 million tax return records in 1988. Table 1 details the approximate number of administrative records that will be used in the 1987 Economic and Agriculture Censuses for the various form types. Clearly, the number of data records received during census years is immense, but the complexity of the required quality assurance goes beyond sheer volume. A data record often contains several data items, each greatly increasing the detail of the individual records and the entire data files. Moreover, not all form types contain the same set of data items, nor do they have the same pattern of receipt. Consequently,

Table 1
The Approximate Number of Administrative Records Used in the 1987 Economic and Agriculture Censuses for the Various Form Types by Tax Year

Type of Record	Number of Records		
	1985	1986	1987
Business Income Tax Files	2,617,000	20,051,000	30,881,000
Form 1040, Schedule C	—	11,750,000	12,500,000
Form 1040, Schedule F	2,450,000	2,450,000	—
Form 1040, Schedule SE	—	—	10,000,000
Form 1120	42,000	2,550,000	2,650,000
Form 1120-A	—	200,000	210,000
Form 1120F	—	11,000	11,000
Form 1120S	17,000	900,000	950,000
Form 1065	108,000	1,750,000	1,800,000
Form 990	—	380,000	400,000
Form 990-PF	—	35,000	35,000
Form 990-T	—	25,000	25,000
Form 1120S, Schedule K-1	—	—	700,000
Form 1065, Schedule K-1	—	—	1,600,000
Annual Tax Files	41,950,000	43,500,000	45,050,000
IRS Business Master File	24,000,000	25,000,000	26,000,000
IRS Payroll and Employment File	17,000,000	17,500,000	18,000,000
SSA Business Birth File	950,000	1,000,000	1,050,000
Total	44,567,000	63,551,000	75,931,000

in addition to performing quality review for over 75 million individual records, the Census Bureau must also be concerned with assuring the quality of the various data items on those 75 million records.

Additionally, businesses file their tax returns with one of ten IRS centers. Each of the individual centers processes the returns, and the quality of data received from different service centers can vary. The Census Bureau reviews data at the service center level in response to such variation.

2. Restrictive budgets

Restrictive budgets are another major factor that contribute to the difficulty of assuring the quality of the administrative record data. In keeping with the overall governmental policy on spending, the Census Bureau is attempting to provide greater services at less cost. Workloads for programming staffs increase significantly during census years, yet the staffs do not expand proportionately. The quality assurance processing, which relies considerably on various computer resources, can be adversely affected. It is also important to note that most quality assurance processing is traditionally done at the mainframe computer levels. Use of the Census Bureau's mainframe computer is costly and becomes more so as increasingly larger data files are processed.

3. Lack of communication between agencies

Miscommunication or lack of communication between agencies has contributed to past administrative record problems. Clear lines of communication between the Census Bureau and the agency providing the data during all phases of the procurement process also are essential for assured data quality. The agencies first must agree upon the data files and the

specific data items that are needed and that can be provided. Certain data that the Census Bureau requests may not be available or in some cases affordable. Any discrepancies must be resolved in time to avoid delays, which could affect data utility. Moreover, the agencies must agree upon the expected quantity and quality of the administrative data. Requirements that quantify the Census Bureau's expectations of the incoming data should be established.

The development and implementation of the quality assurance system represent a comprehensive response to the administrative record data problems we encountered in the past. The system provides for the review of large and complex IRS data files, promotes frequent interagency communication, and identifies errors instantly. The major element of the quality assurance system is the mechanized data verification. Basically, the Census Bureau establishes standards that detail our fundamental requirements of the incoming IRS data. The reporting patterns of the actual data are compared to these standards, and systematic data verification occurs at the microcomputer level. The Census Bureau then prepares status reports indicating whether the data conforms to the standards.

Census Bureau staff members develop the standards far in advance of the actual receipt of the data. This gives the IRS ample opportunity to examine the requirements for reasonableness and request adjustments if necessary. The requirements are divided into timing standards and quality standards. The timing standards list the estimated total number of tax returns for the different types of businesses and the estimated number to be received by various dates. The quality standards detail the expected reporting patterns of specific data items.

The mechanized data verification technique simplifies our analytical review process. A series of results tables are created that compare the actual data to the expected standards. Discrepancy flags are set for those data components that do not meet the standards. This approach minimizes the risk of analytical omissions during the review process.

Status reports comparing the reporting patterns of actual data to the pre-determined standards are sent to the IRS monthly. These status reports are a subset of the comprehensive results tables, detailing only the basic requirements of the IRS data set. The status reports promote communication between the agencies. If data problems exist, they are illustrated in the report. Immediately, the Census Bureau and the IRS must decide upon any remedial action or recovery efforts necessary to prevent compromising the censuses. Timeliness is crucial because the IRS data tapes are not kept indefinitely. If errors are not identified early and remedial action is not implemented in time, recovery of the data may not be possible or may become extremely costly.

The quality assurance system is not designed to guarantee that administrative data problems will never occur. It does serve, however, to document our requirements formally so that the characteristics of the data set are not left to chance, and monitoring and early error identification are possible.

3. DETAILS OF AUTOMATION

Administrative record data files are received weekly and processed initially through mainframe quality assurance programs. The mainframe programs are prepared well before the administrative data files are received and generate the initial quality assurance tables that are fundamental to the entire review process. Traditionally, mainframe programmers were responsible for creating the entire data tables, which included data cells and the surrounding text (*i.e.*, headers and stubs). However, for the data table programs associated with the 1987 Economic Censuses, the two data table components are handled separately. Data tabulation is performed as usual at the mainframe level whereas table text is created at the microcomputer level by non-programmers. A procedure has been developed that generalizes data tables for all administrative

Table 2
Weighted Distribution of Form 1040 Schedule C Records by
Net Receipts Size Class by Service Center

Service Center	Total	Net Receipts Size Class (000)				
		< 0	Blank or 0	1— 2,499	2,500— 4,999	5,000— 9,999
All Centers	1,327,100	200	52,200	149,300	73,900	98,100
Atlanta	133,200	0	5,100	16,500	6,300	11,000
Philadelphia	132,100	100	4,200	11,300	5,300	9,600
Austin	147,600	0	6,300	20,900	9,900	12,900
Cincinnati	153,100	0	5,300	14,900	8,700	9,800
Kansas City	119,500	0	5,500	16,700	7,500	8,500
Andover	111,100	0	3,800	9,800	6,700	8,200
Ogden	162,300	0	7,500	20,200	7,900	11,600
Brookhaven	119,700	0	4,400	12,600	7,100	10,000
Memphis	111,900	100	4,700	14,700	6,700	8,600
Fresno	136,500	0	5,400	11,700	7,800	7,900
Others	100	0	0	0	0	0

Service Center	Net Receipts Size Class (000)					
	10,000— 24,999	25,000— 49,999	50,000— 99,999	100,000— 249,999	250,000— 499,999	500,000 +
All Centers	168,600	185,500	225,100	243,400	87,400	43,400
Atlanta	17,000	19,800	22,200	22,200	8,400	4,700
Philadelphia	17,800	19,800	22,700	27,000	10,100	4,200
Austin	18,700	18,500	22,000	24,900	9,100	4,400
Cincinnati	20,500	20,700	27,300	30,500	9,600	5,800
Kansas City	16,200	15,900	20,700	18,300	6,400	3,800
Andover	13,600	16,700	19,500	20,000	8,800	4,000
Ogden	17,800	19,500	28,800	33,600	11,200	4,200
Brookhaven	16,400	19,700	20,400	19,400	6,400	3,300
Memphis	15,100	14,700	18,600	19,000	6,800	2,900
Fresno	15,500	20,200	22,900	28,400	10,600	6,100
Others	0	0	0	100	0	0

records files. This procedure has allowed the Census Bureau to design a microcomputer program that is capable of building table images for any administrative records file. Once built, the table images are uploaded to the mainframe and used by programmers to align data tabulation files. The job of programming the quality assurance tables is greatly simplified, as table image formation is handled by nonprogrammers, leaving mainframe programmers adequate time to concentrate their efforts solely on data tabulations. Table 2 illustrates one of the various mainframe tables that is produced for each of the different forms of organization. This table shows the weighted distribution of Form 1040, Schedule C records by service center by net receipts size class.

The mainframe computer performs only the basic data tabulations of the administrative records files (*i.e.*, generates current tables). The output from these mainframe quality assurance programs is downloaded to a microcomputer, and all remaining review operations are automated at the microcomputer level. The various operations performed on the microcomputer include calculating percentages used in the review of the current tables, producing

cumulative tables, performing key data item verification, and generating quality assurance status reports. Developing this systematic approach, using mostly micro-computer technology, has allowed greater flexibility of review as well as lessened the workload of mainframe programmers.

The mainframe quality assurance output is imported into a prestructured spreadsheet on the microcomputer. This spreadsheet also will contain the predetermined standards that outline the Census Bureau's expectations of the incoming data set. Automatically, a mechanical table review and data verification are performed; and inconsistencies between the actual data sets and the standards are identified within the results tables. The two major benefits of this data verification system are:

1. It enables us to easily spot problems in the data. Data components that do not meet the standards are flagged for analyst review. The possibility of overlooking errors in the administrative data is minimized.
2. It directs us to areas of the data that require further investigation. The results tables often-times lead us to problems even though the overall standards are met. For example, certain unexpected trends in the results report are reviewed in additional detail. In effect, the results tables enable us to concentrate on those areas that may contain problems. This may involve additional review at the service center level, or it may even require us to download records with these certain characteristics to the microcomputer. We then review these records on a manual basis in an effort to spot the problem.

As previously stated, the standards detail the basic data quality requirements that are essential to the 1987 Economic and Agriculture Censuses. This procedure of automatic quality verification (*i.e.*, comparing the incoming data to predetermined standards) allows us to determine immediately if the basic quality of the incoming data is acceptable.

After current cycle review and verification, cumulative tables are prepared on the microcomputer. This technique of producing cumulative tables on the microcomputer rather than the mainframe provides a more efficient use of our resources. First, it eliminates the need to retain cumulative files on the mainframe system, which reduces computer costs. In the past, these cumulative files were retained on the mainframe and added to each subsequent current cycle to form the next set of cumulative tables. Using microcomputers, simple formulas were established within the spreadsheet that created cumulative tables at virtually no cost. Secondly, the quality assurance tables for the cumulative portion do not require mainframe programming. A printout of the cumulative quality assurance tables are produced and retained for analysis and documentation purposes.

In addition to this comprehensive set of cumulative tables, we produce a set of results tables. As was the case with the current cycle, these results tables detail comparisons of certain key data items. Table 3 shows one of the many results tables that is produced for the cumulative quality assurance. This table details the actual number and percent of the weighted Form 1040, Schedule F records by service center, together with the expected percent. As can be seen, the cumulative data are reasonable and fall within the acceptable standards. If inconsistencies did exist, the applicable service center would have been flagged. The final component of the automated quality review process is the generation of a report detailing the status of the cumulative IRS data file. This report compares the overall quality of the data set to the expected quality indicated in the timing and quality standards. The reports are generated and provided to the IRS approximately monthly. As discussed earlier, the status reports capsulize the quality of the administrative data for representatives of both agencies, which promote frequent interagency communication.

4. RESULTS OF QUALITY ASSURANCE REVIEW

The timing and quality status reports can serve to alert both the Census Bureau and the IRS of data problems in their early stages and facilitate cooperative action by both agencies. In most of the cases, however, the timing and quality standards alert us of changes in respondent reporting patterns. These circumstances require no corrective action by the IRS, but they may have cost and processing implications for the Census Bureau in the 1987 Economic and Agriculture Censuses. Tables 4a and 4b illustrates this point well. Through late May 1987, the Census Bureau had received approximately 697,600 Form 1120 returns (*i.e.*, corporations) with a standard of 760,000 returns. The standard for the number of Form 1120 returns was not met. However, the shortfall in the number of Form 1120 returns was offset by an increase in the number of Form 1120S returns (*i.e.*, S corporations). The Census Bureau had received approximately 328,850 Form 1120S returns, far exceeding the standard of 225,000. The shift in the number of returns for these two types of corporations resulted from the perceived advantages in the new tax law associated with filing Form 1120S rather than Form 1120. Although this represented a legitimate shift in taxpayer reporting patterns that was not a data error, the information was pertinent to our processing. We are implementing a procedure for 1987 that will account for such a shift from corporations to S corporations. Table 5 illustrates one of the various tables from the quality portion of the report. As indicated, the quality of these data meets the standards for each of the basic data items. If an item had failed the standard, it would have been flagged for analyst research.

Table 3
Percent of Weighted 1986 Form 1040, Schedule F Records by Service Center

Tax Year	Total Schedules	Service Centers					
		Atlanta	Philadelphia	Austin	Cincinnati	Kansas City	
1986	Count	2,087,200	176,700	71,600	374,900	262,100	358,600
	Percent	100.0	8.5	3.4	18.0	12.6	17.2
Expected							
	Percent	100.0	8.5	3.0	18.5	11.5	17.5
Expectation ¹ Not Satisfied							

Tax Year		Service Centers					
		Andover	Ogden	Brookhaven	Memphis	Fresno	Others
1986	Count	118,800	343,200	40,300	288,100	52,500	400
	Percent	5.7	16.4	1.9	13.8	2.5	0.0
Expected							
	Percent	5.5	16.5	2.0	14.0	2.5	0.0
Expectation ¹ Not Satisfied							

¹ Acceptance interval of + or - 2.0 percent.

Table 4a
The Weighted Number of 1986 Form 1120 Returns by Various Dates

Date	Form 1120 Returns		Requirement Not Satisfied
	Actual	Required	
Late March 1987	326,500	303,000	Not Satisfied
Late April 1987	697,600	760,000	
Late May 1987		988,000	
Late June 1987		1,190,000	
Late July 1987		1,418,000	
Late August 1987		1,621,000	
Late January 1988		2,077,000	
Late October 1988		2,533,000	

Table 4b
The Weighted Number of 1986 Form 1120S Returns by Various Dates

Date	Form 1120S Returns		Requirement Not Satisfied
	Actual	Required	
Late March 1987	103,350	90,000	
Late April 1987	328,850	225,000	
Late May 1987		292,000	
Late June 1987		352,000	
Late July 1987		420,000	
Late August 1987		480,000	
Late January 1988		615,000	
Late October 1988		750,000	

The automated quality assurance of administrative records files will be completely operational for the 1987 IRS data files. Prototypes of the system have been and are being used for the 1985 and 1986 IRS business income tax files. For both years the automated process and the entire quality assurance system have been instrumental in the successful procurement and review of the IRS data files received for the censuses.

The integration of both mainframe and microcomputer technology in the automated quality assurance system has allowed the Census Bureau to effectively and comprehensively assure the quality of the large data files provided by the IRS. In addition, mainframe computer programmer workloads have been and will continue to be lessened since much of the automation was designed and is controlled by nonprogramming staff and is implemented in a microcomputer environment. Mainframe computer resources are reduced and programming burden is lessened allowing programmers to concentrate their efforts on basic data tabulation. Also important, the automated system provides the flexibility of review for different levels of personnel. Managers can review the summarized timing and quality report and determine the status of the business income tax files quickly and efficiently. Subject-matter analysts will review the more comprehensive quality assurance reports that are produced weekly. As mentioned above, the quality assurance system will direct the analysts to the data elements that require further investigation.

Table 5
Data Element Reporting Patterns of Weighted 1986 Form 1120S Returns

Data Elements	Percent of Form 1120S Returns		Requirement Not Satisfied
	Actual	Required	
EIN			
Blanks, all zeros, or nonnumerics	0.0	Less than 1.0	
Invalid IRD	0.0	Less than 1.0	
PBA CODE			
Blanks or nonnumerics	0.0	Less than 6.0	
Blanks, nonnumerics, unclassified, or invalid PBA codes	11.5	Less than 18.0	
GROSS RECEIPTS OR SALES LESS RETURNS AND ALLOWANCES			
Blanks, all zeros, or nonnumerics	20.9	Less than 40.0	
Of records with a positive numeric entry, the percent in various size ranges:			
- Less than \$100,000	45.7	30.0 — 60.0	
- Greater than or equal to \$100,000 and less than \$500,000	36.9	20.0 — 50.0	
- Greater than or equal to \$500,000	17.4	10.0 — 30.0	
ACCOUNTING PERIOD			
Blanks, all zeros, or nonnumerics	0.0	Less than 1.0	

5. SUMMARY

The Census Bureau has designed an overall quality assurance system that is comprehensive and responsive to the potential problems and limiting factors of complete quality assurance. The system responds to the large volumes of IRS data by interacting with the IRS closely and promptly to ensure proper data procurement. The expected quality of these large data files is jointly determined and agreed upon with the IRS through the timing and quality standards and is verified by the automated QC process. Given this automated process, data verification can occur within the bounds of restrictive budgets and limited programming resources. Microcomputer technology has increased the role and flexibility of subject-matter analysts while lessening the burden of mainframe programmers. Communication with the IRS is frequent and productive, resulting in efficient procurement procedures and improved data quality awareness on the part of IRS and the Census Bureau as well. This collective response to past difficulties will ensure the Census Bureau of receiving the data necessary to conduct the 1987 Economic and Agriculture Censuses in the best manner possible.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments and suggestions.

Using Administrative Record Data to Evaluate the Quality of Survey Estimates

JEFFREY C. MOORE and KENT H. MARQUIS¹

ABSTRACT

The Survey of Income and Program Participation (SIPP) is a new Census Bureau panel survey designed to provide data on the economic situation of persons and families in the United States. The basic datum of SIPP is monthly income, which is reported for each month of the four-month reference period preceding the interview month. The SIPP Record Check Study uses administrative record data to estimate the quality of SIPP estimates for a variety of income sources and transfer programs. The project uses computerized record matching to identify SIPP sample persons in four states who are on record as having received payments from any of nine state or Federal programs, and then compares survey-reported dates and amounts of payments with official record values. The paper describes the project in detail and presents some early findings.

KEY WORDS: SIPP; Record check; Record linkage; Survey response validity.

1. INTRODUCTION

This paper addresses issues concerning the use of records to evaluate the quality of survey estimates and describes a specific application to the Survey of Income and Program Participation (SIPP) in the United States.

Matching administrative records to survey observations on a case-by-case basis, which we call a “record check,” provides useful information to survey users and designers. A record check enables the analyst to make a full range of measurement error parameter estimates for evaluation purposes. These estimates, in turn, facilitate two basic kinds of activities:

1. quantifying the effects of measurement errors on subject-matter estimates such as means, proportions, correlation coefficients, and multivariate regression coefficients (and possibly adjusting the estimates to correct for the measurement errors), and
2. deriving more efficient survey designs that directly address, for example, the tradeoffs between measurement quality and costs.

1.1 Basic Terms

Our focus here is on measurement (or “response”) errors, although the record check method can be extended to evaluate other nonsampling and sampling errors also. This is not a technical exposition, but we do need to define some of our basic terms first. We assume that the survey observation from sample element i can be expressed as the sum of the true value and an error, e :

$$\text{Survey}_i = \text{True}_i + e_i.$$

¹ Jeffrey C. Moore and Kent H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, Room 433 Washington Plaza Building, Washington, DC, 20233. This is a revised version of a paper presented at Statistics Canada’s International Symposium on Statistical Uses of Administrative Data, November 23–25, 1987. This paper presents the views of the authors and does not necessarily represent official Census Bureau policy or opinions.

The average bias in a set of N survey observations, which we call the response bias or survey bias, is

$$\bar{e} = \sum e_i / N,$$

and the response error variance is just $\text{Var } e$.

Similarly, the measurement model for the administrative record observation is:

$$\text{Record}_i = \text{True}_i + u_i,$$

so that record bias is \bar{u} and record error variance is $\text{Var } u$.

1.2 Comparison of Evaluation Approaches

The capabilities of the record check approach can be contrasted with other methods of evaluation such as reinterviews and experiments. Reinterviews and other repeated measures designs aim at estimating a very limited set of measurement error parameters, usually something called the simple response variance or the response error variance. These approaches implicitly make strong assumptions about true change over time and about either the true value or bias parameter (Marquis 1986).

One frequently attempted remedy is to create a true value measurement as part of the reinterview program, for example by reconciling discrepant answers with a knowledgeable respondent or by asking much more detailed and specific questions during the reinterview. But the validity of these "true value" measures is suspect. Both Bailer (1968) and Koons (1973) have shown, for example, that reconciled reinterview responses are biased. And while detailed, specific questioning is often preferred to a more global approach, there is no independent evidence that it reduces measurement biases to zero — or at all. Record checks potentially provide higher quality criterion information requiring much weaker (and perhaps more realistic) assumptions for purposes of estimating survey data quality.

A different method of evaluating aspects of surveys is the experiment, such as a fully-crossed factorial design or an interpenetrated design for assigning interviewers. Analysts compare experimental groups with respect to statistics such as subject matter means or proportions and draw conclusions about which treatment produces more or less reporting of the subject matter of interest. What is controversial, however, is determining which is "better" in a measurement sense, a difficulty that is much reduced when criterion data — such as administrative records — are available.

Without criterion data, it is often necessary for the analyst to resort to strong assumptions about measurement errors, such as:

1. more reporting is better reporting;
2. forgetting of meaningful material increases with the passage of time;
3. unbounded interviews contain overreports, bounded interviews don't;
4. reporting performance decays with length of interview or time-in-sample;
5. people are basically lazy and devious — they will lie to avoid being asked a detailed set of questions; and
6. self reports are better than proxy reports.

Indeed, these assumptions have become part of the folklore of survey design in the western world. And yet, it is difficult to find any support for any of these assumptions from appropriately designed record checks. Experiments and related arrangements are excellent approaches to pinpointing the sources of variation, and in untangling estimation problems of collinearity, but are often unnecessary and seldom sufficient for evaluating an existing measurement process.

In sum, these other evaluation approaches are forced to make strong assumptions about:

1. the independence of the original and evaluation measures when they are clearly dependent;
2. the relationship of the original measure to a criterion when no objective, external link exists; and/or
3. cognitive processes not supported by research.

Record checks also employ assumptions in evaluating measurements. For example, the usual way of estimating the response bias is to assume no record bias ($\bar{u} = 0$) and simply calculate the average of the differences between the matched survey and record observed values:

$$\text{Estimated Survey Bias} = \sum (S_i - R_i)/N.$$

While one cannot directly support the no-record-bias assumption, one can conduct meaningful sensitivity tests of the effects of possible violations of the assumption on evaluation conclusions.

1.3 Issues in Designing Record Checks

Several issues merit consideration in designing a record check to evaluate survey measurement. We comment on some of the main ones here: incomplete observation designs, matching errors, record errors, true value differences, and absence of repeated measures or experimental design features.

1.3.1 Incomplete Observation Designs

Past record checks have often used one-directional or partial designs for data collection, such as when we survey people about owning library cards and check the records for those who claim to have one, or sample from a list of people with a diagnosed chronic disease and survey them to see if they report it in a survey questionnaire. Because these partial designs do not observe the full range of response errors in the correct proportions, they yield biased estimates of such classical measurement error parameters as the response bias and the response error variance. One-directional designs can fail to detect some or all of the true survey bias, can cause the analyst to interpret up to one-half of the response error variance as response bias, and can predetermine the sign of the estimated response bias if the measured variable is binary (Marquis 1978). Full designs are a necessary (albeit not sufficient) condition for obtaining unbiased estimates of the desired response errors.

1.3.2 Matching Errors

The essence of the record check is a one-to-one matching of survey and record observations. This is difficult to do correctly, and matching errors (false matches, false nonmatches) will potentially bias the measurement error estimates of interest. Neter *et al.* (1965) show that when there are no unmatched cases, the mismatches will bias the estimates of response error variance upward. In terms of the reliability of a dichotomous measure (which is a function of the response error variance), the estimate will be attenuated by exactly the match error rate (Marquis *et al.* 1986). It is therefore desirable to keep match errors to a minimum and to know something about the errors that remain.

1.3.3 Administrative Record Errors

As noted earlier, one usually has confidence that the records in a record check study are very good measures of the trait of interest. If the implied assumptions about record measurement bias and record measurement error variance are violated, this can cause the response error

estimates to be biased away from zero. For example, bias in the record observations can appear as bias in the survey observations but with the opposite sign. Feather (1972) describes this effect in a record check of physician visits in Saskatchewan, in which an apparently large survey over-reporting rate was due to the record's recording a complete treatment procedure rather than the individual visits for diagnosis. Similarly, the presence of measurement error variance in the record can cause inflated estimates of response error variance in the survey (Marquis 1978).

1.3.4 True Value Differences

Problems arise when the survey and record systems use different definitions. This is often the case in "aggregate comparisons" of population parameter estimates made separately by each source. A common difference is in the scope of the populations covered, such as when the survey frame is limited to the civilian, noninstitutionalized population and the record includes everybody. Case-by-case matching can minimize the threats posed by differential coverage, but even estimates derived from these studies can still be plagued by differences in the concepts or the attributes of the concept. For example, Cox and Iachan (1987) report the results of a study which compared survey-reported health conditions with medical records. The authors conclude that a major reason for the lack of correspondence between survey and record reports was differing concepts — the survey was designed to elicit the complaints which led to doctor visits while the medical records focused on final diagnoses. As an example from our study, the administrative records often contain the date a check was written for a transfer payment, while our survey respondents tell us when they received the payment. Such differences can threaten our time-related estimates of such things as telescoping response errors.

1.3.5 Absence of Experiments and Reinterviews

Evaluation record checks can detect errors but are not good at evaluating the remedies for the errors. To know how well a different survey design might perform, one must usually either test the alternative design options or arrange to estimate parameters of an underlying model from which survey designs can be derived (*e.g.*, a model of forgetting effects). For example, an evaluation record check design can estimate and compare response errors for self and proxy respondents. Without heroic assumptions it cannot, however, suggest how the measurement error parameters would change if the survey's respondent rule were changed (say, to allow only self response).

Similarly, a record check without a reinterview or another set of independent measures is limited in the number of basic error parameters it can estimate. For example, our initial definitions mentioned three parameters: true value, survey error, and record error. Without a reinterview (or other independent measure) there are only two measures with which to estimate the three unknowns. An additional measure can help identify the estimates of the parameters in the model.

2. CHARACTERISTICS OF SIPP

Here we briefly describe the main features of SIPP — the Survey of Income and Program Participation — as a prelude to discussing the record check evaluation design.

2.1 Overview of SIPP Contents

The purpose of SIPP is to provide improved information on the economic situation of people and households in the United States. It collects comprehensive longitudinal data on cash and noncash income, eligibility for and participation in Government transfer programs, assets and

liabilities, labor force participation, and a host of related topics. SIPP data assist the evaluation of the cost and effectiveness of current Federal government programs, the potential impacts of proposed program changes, and the actual impacts of changes when implemented. In general, the Census Bureau and other Government agencies which have fostered and supported the development of SIPP expect it to be an invaluable tool for domestic policy planning (Nelson *et al.* 1985).

Core SIPP questions — repeated in each wave of interviewing — cover labor force participation and amounts and types of income received, including transfer payments and noncash benefits from various programs for each month of the reference period. The core questions cover nearly 50 sources of income, including Government transfer payments from retirement, disability and unemployment benefits, and welfare programs such as Aid to Families with Dependent Children. Information is also gathered on noncash programs such as food stamps, Medicare and Medicaid; private transfers such as pensions from employers, alimony, and child support; ownership of assets that produce income, such as interest, dividends, rent and royalties; and on miscellaneous sources of income, such as estates.

2.2 SIPP Data Collection Design

SIPP started in October 1983 with a sample of approximately 25,000 designated housing units (the “1984 Panel”) selected to represent the noninstitutional population of the United States. In February 1985 a new and slightly smaller panel was introduced. Additional panels are to be introduced each February throughout the life of the survey. Due to budget reductions, the sample size for new panels is currently about 15,000 households.

Each sample household is interviewed by personal visit once every four months for 2-1/2 years, resulting in a total of eight interviews. The reference period for each interview is the four months preceding the interview month. At each visit to the household, each person fifteen years of age or older is asked to provide information about himself/herself. Proxy reporting is permitted for household members not available at the time of the visit. Information concerning proxy response situations is recorded and is available for analytical purposes.

To facilitate field operations, each sample panel is divided into four subsamples (“rotation groups”) of approximately equal size, one of which is interviewed each month. Thus, one “wave” or cycle of interviewing is conducted over a period of four months for each panel. This design produces steady field and processing workloads, but it also means that each rotation group uses a slightly different four-month reference period.

Beginning with the second wave of interviewing in the 1984 panel, SIPP conducts reinterviews with a small sample of households about a subset of items (including program participation). These data are used to check for interviewer falsifications and perhaps to estimate response inconsistencies.

3. RECORD CHECK DESIGN

The purpose of the record check is to provide an evaluation of some of the income data gathered in SIPP. We highlight important features of the design of the record check next, covering the samples, the administrative records, the matching approach, and the analysis.

3.1 Record Check Samples

The SIPP Record Check uses a “full” rather than a one-directional design; that is, the records allow us to validate all observed values in the survey. Design options we did not choose include:

1. checking records only for people who claimed to be participating in a program, or
2. drawing a sample of known recipients and interviewing them to determine how truthfully they report.

Both of these designs are incomplete and will result in biased estimates of the response error parameters.

The Record Check Study restricts attention to a subset of available SIPP data from the 1984 Panel. First, the sample of people is restricted to households in four target states: Florida, New York, Pennsylvania, and Wisconsin. In the 1984 Panel this translates to approximately 5,000 households. Second, the study's sample of time periods includes only the first two waves of the 1984 Panel. Figure 1 illustrates the wave, rotation group, interview month, and reference period structure for the target survey data.

Third, the SIPP Record Check Study focuses on the quality of reciprocity and amount reporting for selected Government transfer programs. It compares survey reports and administrative records for five Federally-administered programs (Federal Civil Service Retirement, Pell Grants, Social Security (OASDI), Supplemental Security Income (SSI), and Veterans' Compensation and Pensions), and four state-administered programs (Aid to Families with Dependent Children (AFDC), food stamps, unemployment compensation, and worker's compensation).

We limited the study to four states — Florida, New York, Pennsylvania, and Wisconsin — in order to keep the study to manageable proportions. Major criteria used to select these states were:

1. the presence of a computerized, accessible, and complete record system for all target programs;
2. a large SIPP sample;
3. reasonable geographic diversity; and
4. a willingness to share individual-level data for purposes of this research.

Thus, the states were selected purposively; no attempt was made to sample states to be representative of the Nation.

We requested from each participating state agency identifying and receipt information for all persons who received income from the target program at any time from May 1983 through June 1984. The identical request was made of the participating Federal agencies, with the exception that only recipients residing in one of the four selected states were to be included in the data extract.

Wave	Rotation Group	Interview Month	Reference Period Months											
			Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	
1	1	Oct 83	X	X	X	X								
	2	Nov 83		X	X	X	X							
	3	Dec 83			X	X	X	X						
	4	Jan 84				X	X	X	X					
2	1	Feb 84					X	X	X	X				
	2	Mar 84						X	X	X	X			
	3	Apr 84							X	X	X	X		
	4 ¹	May 84								X	X	X	X	

Figure 1. Survey Structure for Data Included in the SIPP Record Check Study.

¹ Technically, rotation group 4 of the 1984 SIPP Panel was not administered a Wave 2 interview. The "missing" interview was transparent to respondents, however, who were simply given their Wave 3 interview at the time they would have received the Wave 2 interview. For present purposes, the Wave 3 interview for rotation group 4 is identical to the Wave 2 interview for all other rotation groups, and is included in the Record Check Study in order to have two interviews from all sample cases. All references in the text of this paper to "Wave 2" include the Wave 3 interview for this portion of the panel.

As noted earlier, errors in the records can cause problems for record check evaluation studies. Although several of the administrative record files obtained for this project contain very minor deficiencies, only two appear likely to pose major analytical problems: the New York worker's compensation file, and the Veterans' Compensation and Pensions file. Each is known to be incomplete in its coverage of recipients. The New York file excludes an unknown number of cases which were "closed" (*i.e.*, cases which had already been adjudicated and for which payments by a private insurance carrier had already begun) at the time the data base was created several years ago. The Veteran's file excludes the approximately one percent of all recipients whose benefits were sent to a financial or other institution. There are no known coverage problems with any other files.

An unavoidable problem which afflicts all of the administrative files to some extent is the discrepancy between payout date and receipt of payment; obviously, the SIPP respondent reports the latter and has no knowledge of the former, and the reverse is true for the program records. Where the payout date is close to the end of a month it may be difficult to distinguish a forward telescoping error from a legitimate difference between month of payment and month of receipt. Where there are definitional discrepancies, such as this payment date issue, our analyses will attempt to model them explicitly.

4. MATCHING

4.1 Introduction

The quality of matching has important effects on some of the most critical response error estimates, such as the response error variance. Ideally, variables used to match survey and record observations are measured without error and are able to identify an individual uniquely. The ideal, of course, is never realized.

However, the variables we have available to match surveys and records should go a long way toward minimizing the match errors. Some, such as social security number (SSN), uniquely identify an individual even if other information such as address is outdated, garbled, or obliterated or missing. For purposes not directly related to this study (although certainly of benefit to it), the Census Bureau has taken special measures to ensure that SSN information as reported to the SIPP is complete and valid. For all Wave 1 and 2 sample persons, reported SSN's and reports of not having an SSN were verified and, if necessary, corrected, by the Social Security Administration. Sater (1986) estimates that as a result of this operation the SIPP file contains a valid SSN for about 95 percent of SIPP sample persons who have one.

The wealth of other data — last name, first name, house number, street name, apartment designation, city, zip code, sex, and date of birth — is sufficient for high quality matching even in the absence of a unique identifier such as SSN. In addition, to aid us in evaluating the impact of any remaining match errors, the Census Bureau's matcher produces an ordinal measure of the goodness of the match/nonmatch of each survey observation to its appropriate administrative record counterpart.

4.2 The Census Bureau's Computerized Match Procedures

The Record Check Study uses computerized matching procedures applying the theoretical record linkage work of Fellegi and Sunter (1969). The process involves multiple discrete steps, but basically there are four:

1. standardizing the common data fields in the two files which the matcher will examine to determine whether a pair of records is a match or not;
2. sorting the two files into small subsets of records (or "blocks") which constitute a feasible number of pairs to be examined by the matcher;

3. determining and quantifying the usefulness of each data field to be considered in the match for identifying true matched pairs; and
4. implementing the computer algorithms which perform the actual record matching.

4.2.1 Standardization

The Record Check Study processes all data files — both the SIPP files and the administrative record files — through an address standardizer which standardizes the format of various components of an address (*e.g.*, street name, type, and direction; city name; state abbreviation; *etc.*) and parses each component into a fixed data field. Several programs have been developed for this purpose; we use the ZIPSTAN standardizer developed at the Census Bureau.

In addition to the standardization procedures which apply to all data files, many files require modifications to individual data fields to ensure a common format across files for matching. Common examples of variables which pose problems of this type are sex (which can be represented by either an alpha (“m” or “f”) or a numeric (“1” or “2”) code); date of birth (which has many variants — *e.g.*, “mm-dd-yy,” or “cc-yy-mm-dd,” or the Julian format); and name (which may be a single field or which may have separate fields for each component). We prepare custom-made programs for this type of standardization.

4.2.2 Blocking

Blocking — establishing subsets of records for the matcher to examine in searching for matched pairs of records (*e.g.*, Jaro 1985) — is necessary when matching files with large numbers of records. Obviously, the probability of finding all true matches would be highest if, for each record on one file, the entire other file were searched for a match. However, for large files such unrestricted searches for matched records are simply not feasible. Blocking each file into subsets of records makes matching large files feasible, but at the cost of excluding some records from the search; it thus increases the likelihood that some true matches will be missed. Ideal blocking components, therefore, have sufficient variation to ensure the partitioning of the files into many (and therefore smaller) blocks, and are effective match discriminators — that is, nearly always agree in true match record pairs and nearly always disagree in true nonmatch record pairs.

The study uses multiple independent blocking strategies for each pair of files to be matched, thus minimizing the likelihood that a true match pair will escape detection as a result of blocking. One primary blocking strategy employs the first three digits of the United States Postal Service’s five-digit ZIP code and a four-character SOUNDEX code derived from the sample person’s/recipient’s last name. The ZIP code is a sub-state geographic indicator which generally is recorded quite accurately according to Census Bureau matching experts. The SOUNDEX algorithm is widely-used for creating a standard length, standard format code from input character strings of varying lengths; its advantage for blocking purposes is that it minimizes blocking errors due to misspellings, although it cannot eliminate such errors entirely. The second primary blocking arrangement uses the last four digits of the SSN.

4.2.3 Data Field Match Weights

With some variation, the data fields used in the matching of the SIPP and administrative record files include house number, street name, apartment number, city, ZIP code, SSN, sex, date of birth, last name, and first name. Intuitively, these fields are not equally useful in determining whether a particular pair is a match or not — as an obvious example, agreement on sex is not as indicative of a true match as is agreement on SSN. Fellegi and Sunter (1969) include, in their presentation of a general theory of record linkage, discussions of weight calculations

reflecting different data fields' differing discriminating powers and how these weights feed into optimal decision rules. The Census Bureau's Record Linkage Research Staff has developed programs using Newton's method for non-linear systems (see Luenberger 1984) to solve the Fellegi-Sunter equations, and these programs are used in the SIPP Record Check Study to compute final match weights.

4.2.4 The Computer Matcher

The Census Bureau's computer matcher executes the Fellegi-Sunter procedures on a user-defined set of data fields on files sorted (blocked) according to user specifications. For each data field to be considered in the match, the user supplies match weight seed values, defines the type of agree/disagree comparison (whether the fields must be exactly comparable in order for the matcher to treat them as agreeing, or whether only approximate comparability is necessary), and identifies missing value entries and specifies how they are to be treated (included or ignored in the calculation for a composite match weight). The user sets the composite weight cutoff values for matched pairs and nonmatched pairs, and generates the appropriate COBOL program codes to conduct a match through GENLINK, the Census Bureau's Record Linkage Program Generator (LaPlant 1987).

In simple terms, the matcher:

1. searches each data file for comparable blocks of records — that is, records which agree exactly on the designated blocking components;
2. counts the number of records in found blocks to ensure that neither file's block size exceeds the preset maximum;
3. computes a composite match weight for all possible pairs of records in the block;
4. within the block, assigns each record in one file to a paired record in the other file according to a formula which maximizes the total composite weight for all pairs in the block;
5. applies the Fellegi-Sunter decision procedure to determine whether a pair is a match, a nonmatch, or requires further review; and
6. produces a "pointer" file map to the paired records in each file.

5. ANALYSIS

Our goals for the record check study are to estimate selected measurement error parameters for our samples of people, content, and times, and to assess how these errors relate both to each other and to variables that reflect survey design features. Our general plan is to use the matched data to estimate for each dichotomous participation variable:

1. the response bias (using the survey-minus-record difference score);
2. predictors of the response bias (using logistic or probit regression techniques or possibly LISREL techniques based upon matrices containing polyserial and tetrachoric coefficients of association (Jöreskog and Sörbom 1984);
3. the response error variance (*e.g.*, derived from regression residuals);
4. the conditions or groups associated with very large and very small response error variances; and
5. the kinds and amounts of confusion among transfer programs that contribute to the response errors (using covariance structure analysis procedures such as LISREL).

(We will estimate the same parameters for reports of the amounts of money received from each transfer program but have not yet selected our basic estimation approach.)

The measurement error issues to be addressed fall into one of two categories: issues which apply to all time periods and issues that require comparing errors across time periods. In the former category are estimates of the amounts of response errors for self and proxy respondents or contributed by interviewers. In the latter category are the errors arising from panel surveys with familiar labels such as telescoping, time-in-sample bias, memory decay, rotation group bias, *etc.* — those implying that measurement errors will differ across time periods when everything else is held constant. To this list we add what Hill (1987) has referred to as the “seam” bias in longitudinal surveys, which we discuss below.

To appreciate the applied questions we wish to address about the different time periods, consider Figure 2, which presents the interview and reference month calendar for one rotation group of SIPP respondents:

The figure shows two interviews. The first takes place in early October and asks about what happened in September (last month), August (two months ago), July (three months ago), and June (four months ago). Similarly, the second interview, taken four months later, asks about January, December, November, and October. We refer to the transition between September and October as the “seam” because it is between the reference periods covered by the two interviews.

To investigate the internal telescoping hypothesis (which asserts that events are not forgotten, just remembered as having happened closer to the present time), we will be testing whether the response bias for the early months of the reference period (June and July in Wave 1 and October and November in Wave 2) is negative and the response bias for later months (August and September or December and January) is positive, and that the two biases sum to zero.

We plan to test the bounded interview hypothesis, which says that events from the remote past are reported as happening within an unbounded reference period (June through September), but that this will not happen in reference periods bounded by a previous interview (here, October through January).

To examine the hypothesis about memory decay (that the probability of forgetting an event increases with the passage of time), we will test whether the response bias is more negative for the early months of each reference period than for later months.

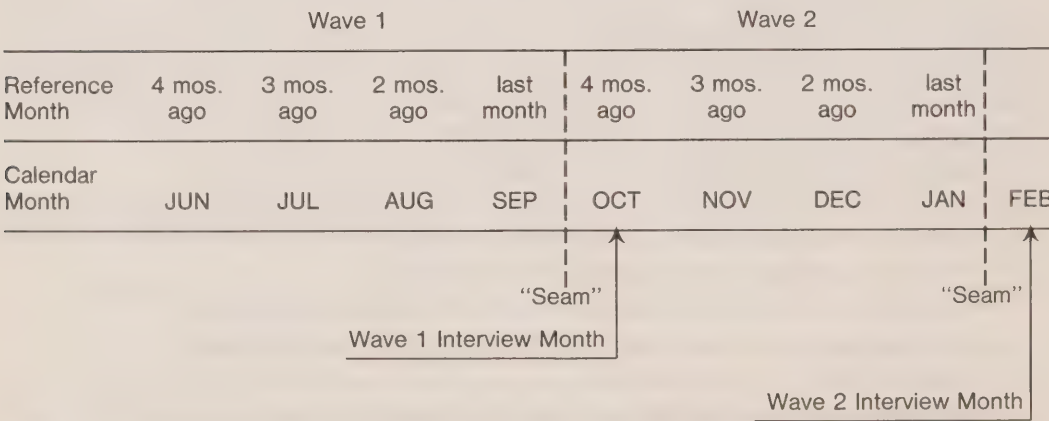


Figure 2. SIPP Survey Time Periods for Rotation Group 1 Showing Reference Months, Calendar Months, Interview Months, and Interview “Seam”.

The time-in-sample and rotation group hypotheses suggest that response errors will be greater in the second interview than the first, after correcting for any seasonal effects. We plan to examine this and, if we find it to be true, test some of the ideas in the literature about why it may be true. Are the sample elements that survive from the first to the second interview different, as Stasny and Feinberg (1985) suggest, or does the quality of the survivors' reporting deteriorate, as the Neter and Waksberg (1966) conditioning hypothesis might predict?

We don't know yet the extent to which SIPP is experiencing these more traditional problems of longitudinal surveys. One problem for which there is evidence, however, concerns the estimation of month-to-month changes in program participation (Burkhead and Coder 1985). Specifically, more changes in program participation take place at the "seam" between interviews (between September and October in Figure 2) than between the months covered by any one interview (*e.g.*, between June and July or July and August or August and September). The Census Bureau has not published monthly program participation transition estimates from SIPP yet because the estimates show a pattern that appears to be affected heavily by measurement error. Moore and Kasprzyk (1984) and Hill (1987) have speculated about what kinds of response, nonresponse, or procedural errors might be producing the pattern and which set of transition estimates is more accurate. By addressing the problem with administrative data, we hope to come much closer to a definitive explanation about the role of response and nonresponse errors in producing the observed pattern.

Related, possibly, to the seam bias issue is the better-understood phenomenon that measurement error variance tends to inflate estimates of gross change or underestimate stability. Recent literature (*e.g.*, Fuller and Tin 1986) suggests several possible approaches to the problem. We plan to begin the empirical exploration of the measurement error effects on the transition estimates to learn whether, for example, we can base corrections for the response errors on estimates from reinterviews.

Finally, we have hinted previously at the problems that may arise in getting unbiased estimates of the errors if the records also contain errors. We plan, with the use of reinterview measures (that identify the estimate of $\text{Var } e$) to estimate the record error variance ($\text{Var } u$). However, we have no plans to relax the assumption that the records are unbiased.

6. PRELIMINARY FINDINGS

To illustrate our approach, we examine the "seam" issue with data for two Government transfer programs in one state. Recall that the seam problem is that monthly survey reports about program participation status show more frequent status changes between months covered by separate interviews than between other months (covered by the same interview). With the administrative record data we are able to begin to answer key questions concerning the quality of SIPP transition estimates: Are too many transitions reported at the seam? Are too few reported for other months? Does SIPP capture the right number of changes over the whole reference period but distribute them incorrectly?

Figures 3 and 4 contain results of our initial seam bias analyses. Data for these initial analyses come from matched/merged SIPP and administrative record files for Aid to Families with Dependent Children (AFDC) and food stamps in the state of Wisconsin.

A total of 1,632 people were eligible SIPP sample persons in Wisconsin in Wave 1 of the 1984 SIPP Panel. Of this total, 92 (6%) refused to report an SSN and were excluded both from the administrative record match and from the response error analyses. Also, the sample residing in Wisconsin is part of a national sample and is not necessarily representative of Wisconsin.

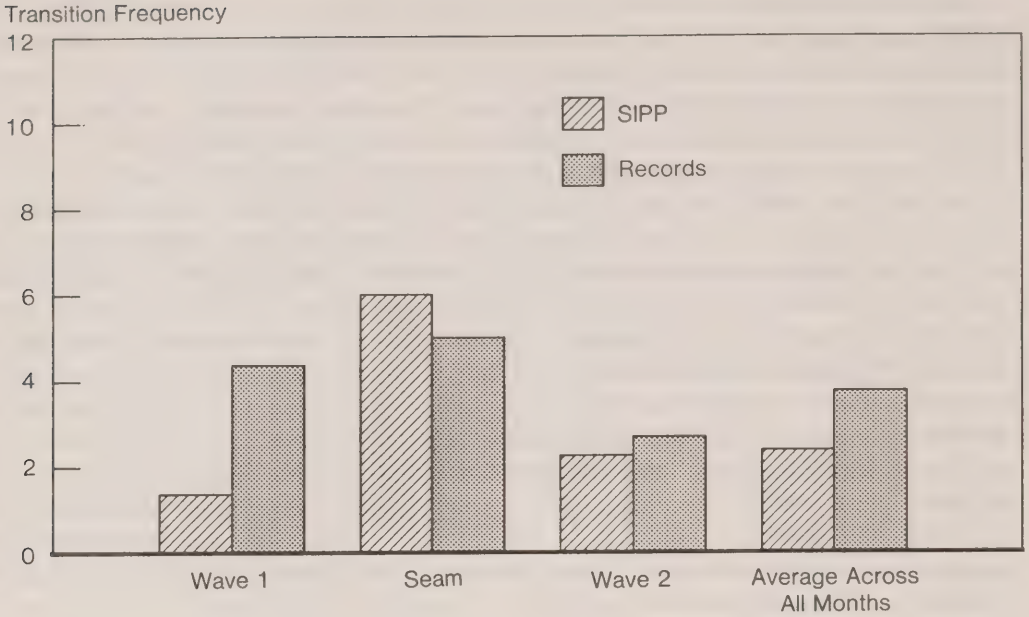


Figure 3. Month-to-Month AFDC Participation Transitions: Comparison of Transition Frequency at the Seam with the Average Frequency Within Waves 1 and 2, and the Overall Average Across All Months, for SIPP and Administrative Records.

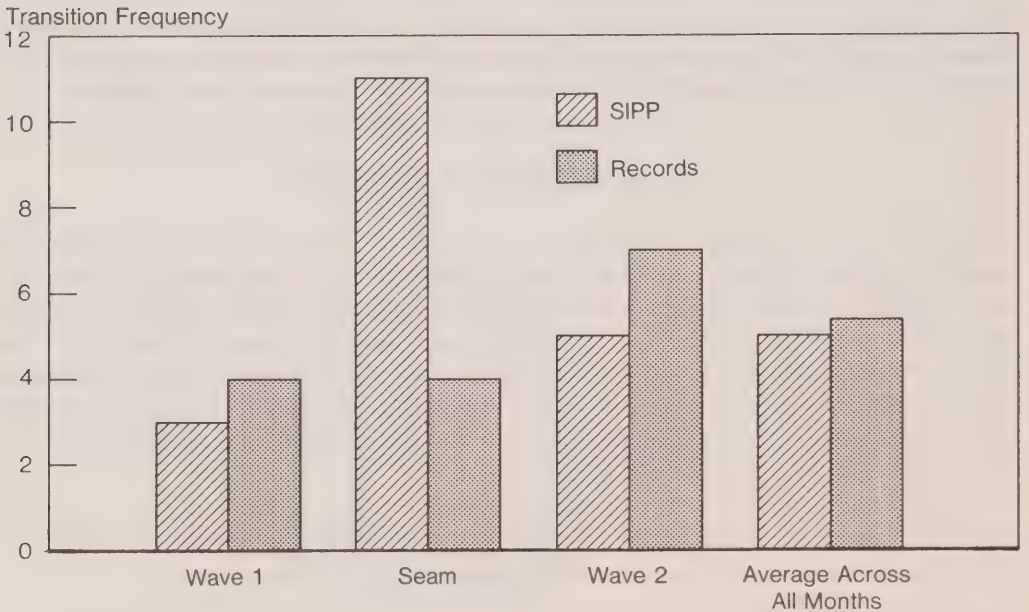


Figure 4. Month-to-Month Food Stamps Participation Transitions: Comparison of Transition Frequency at the Seam with the Average Frequency Within Waves 1 and 2, and the Overall Average Across All Months, for SIPP and Administrative Records.

SIPP procedures assume that all sample persons identified in Wave 1 were eligible sample persons in the same household for all months of the Wave 1 reference period, and that no one other than those eligible at the Wave 1 interview was a household member in the preceding four months. Thus, the month-to-month transition estimates within Wave 1 derive from a constant respondent base of $(1,632-92 =) 1,540$ people. In Wave 2, however, the fluidity of household composition is recognized, resulting in respondent bases which vary slightly from one month-pair to the next — including the interview seam. In the data below the number of eligible persons in both “seam” months is 1,517; within Wave 2 the respondent bases for the three month-pairs are 1,522, 1,531, and 1,532. (Separate analyses (not shown here) indicate that the trends shown in Figures 3 and 4 are not sensitive to excluding people not present in all eight months of the Wave 1 and 2 reference periods.) Because of the small number of cases and the unrepresentative nature of the Wisconsin sample we do not offer inferential statistics for this set of illustrations.

In the figures, the striped bars indicate the number of transitions according to administrative records and the empty bars indicate the number of transitions according to SIPP. If there are too many SIPP transitions at the seam, the empty bar should tower over the striped bar for the comparisons labelled “Seam.” If there are too few transitions reported in SIPP for the months covered within an interview, the empty bar should be smaller than the striped bar for the comparisons labelled “Wave 1” and “Wave 2.” And, if SIPP interviews yield approximately the right number of transition reports, the empty and striped bars should be approximately the same height for the comparisons labelled “Average Across All Months.”

Figure 3 presents the average frequency of month-to-month transitions in Wisconsin AFDC participation within Waves 1 and 2 for the two data sources, and contrasts those figures with the number observed at the Wave 1/2 interview seam. The SIPP “seam bias” problem is quite apparent — the frequency of transitions at the seam is greater than the average within either interview. Although the absolute differences with this sample size are small, the record data suggest that the AFDC seam bias results from a combination of too many transitions reported at the seam and too few in the within-interview months. The final columns of Figure 3 suggest, additionally, a net underreporting of AFDC transitions in SIPP, in addition to the time placement problem.

The Wisconsin food stamps results are summarized in Figure 4, where the seam bias effect in SIPP is even clearer. Once again, the administrative record data suggest a tendency for within-interview transitions to be consistently underestimated with SIPP data. And, in this instance the contrast of survey and record data is even more clear in indicating that SIPP seam transitions are severely overestimated. Unlike the AFDC results, however, both survey and record contain about the same number of transitions overall, suggesting just a time placement problem and not a net underreporting bias.

7. CONCLUSIONS

After a lengthy matching and file preparation process, we are just beginning our analysis of this rich data set. However, with just the initial results presented here we have already shown how record check findings can contribute to our understanding of important measurement error issues — in this case, the SIPP seam bias. There are many more tests to be done and many hypotheses to explore before we can draw definitive conclusions about the nature of SIPP measurement errors and their probable causes. We are confident that the SIPP Record Check Study will allow us to make important advances toward understanding the sizes and forms of these survey errors and perhaps suggest their causes.

ACKNOWLEDGEMENTS

The SIPP Record Check Study has already benefited greatly from the efforts of many people. While we cannot list here all who deserve recognition, we do gratefully acknowledge the particular contributions of: Jeannette Robinson, for preparing the multitude of administrative record files for matching; Elaine Fansler, for preparing and executing countless computer match runs; Bill LaPlant, for sharing his considerable expertise regarding the Census Bureau matcher and attendant software; Chris Dyke, for his efforts to assist in making the matching system work on a new computer system; and Dan Kasprzyk, for his constant and patient support of this entire endeavor. This paper has also benefited from thoughtful reviews by the editor and two anonymous Survey Methodology Journal referees, whose comments and suggestions we also acknowledge.

REFERENCES

- BAILAR, B. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.
- BURKHEAD, D., and CODER, J. (1985). Gross changes in income reciprocity from the Survey of Income and Program Participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-356.
- COX, B., and IACHAN, R. (1987) A comparison of household and provider reports of medical conditions. *Journal of the American Statistical Association*, 82, 1013-1018.
- DAVID, M. (1983). *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program*. New York: Social Science Research Council.
- FEATHER, J. (1972). *A Response Record Discrepancy Study*. Saskatoon, Saskatchewan: University of Saskatchewan.
- FELLEGI, I., and SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- FULLER, W., and TIN, C. C. (1986). Response error models for changes in multinomial variables. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, 425-441.
- HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 210-215.
- JARO, M. (1985). Current record linkage research. Paper presented to the Census Advisory Committee of the American Statistical Association, U.S. Bureau of the Census, April 25.
- JÖRESKOG, K., and SÖRBOM, D. (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*. Mooresville, Indiana: Scientific Software, Inc.
- KOONS, D. (1973). Quality control and measurement of nonsampling error in the Health Interview Survey. *Vital and Health Statistics*, Series 2, 54.
- LAPLANT, W. (1987). Maintenance Manual for the Generalized Record Linkage Program Generator (GENLINK) SRD Program Generator System. Statistical Research Division Internal Working Paper, U.S. Bureau of the Census.
- LUENBERGER, D. (1984). *Linear and Nonlinear Programming*. Reading, Massachusetts: Addison Wesley.
- MARQUIS, K. (1986). Discussion of 'Correlates of reinterview inconsistency in the Current Population Survey'. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, 235-240.

- MARQUIS, K. (1978). Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations. Technical Report R-2319-HEW, The Rand Corporation.
- MARQUIS, K., MARQUIS, S., and POLICH, M. (1986). Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association*, 81, 381-389.
- MOORE, J., and KASPRZYK, D. (1984). Month-to-month reciprocity turnover in the ISDP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 726-731.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). An Overview of the Survey of Income and Program Participation, Update 1. SIPP Working Paper Series, No. 8401, U.S. Bureau of the Census.
- NETER, J., MAYNES, S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- NETER, J., and WAKSBERG, J. (1966). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- SATER, D. (1986). SSN Response Rates and Results of SSN Validation/ Improvement Operation. Internal memorandum, U.S. Bureau of the Census.
- STASNY, E., and FIENBERG, S. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in the Labor Force Statistics*, U.S. Departments of Commerce and Labor, 25-39.

The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities

COLLEEN CLARK and ROBERT LUSSIER¹

ABSTRACT

Statistics Canada is currently rebuilding its central register of economic entities. The new register views each economic entity as a network of legal and operating entities whose characteristics allow for the delineation of statistical entities. This network view, the profile, is determined through the 'profiling' process which involves contact with the economic entity. In 1986 a list of all entities in-scope for a profiling contact was required so that profiles could be obtained to initialize the new register. Administrative data were used to build this list. In the future, administrative data will be a source of information on changes that may have happened to economic entities. They may thus be used as a source of direct update or as a signal that a review of the structure of an entity is required. The paper begins with the objectives of the profiling process. The procedures for constructing the frame for the initial profiling process using several administrative data sources are then presented. These procedures include the application of concepts, the detection of overlap between sources, and the evaluation of data quality. Next, the role of administrative data in providing information on changes to business entities and in requesting profiles to be verified is presented. Then the results of a simulation study done to assess this role are reviewed. Finally, the paper concludes with a series of questions on the methodology of using administrative data to maintain profiles.

KEY WORDS: Administrative data; Central register; Profile.

1. INTRODUCTION

Statistics Canada is in the process of reorganizing its programme of economic surveys. The new programme will result in an increased use of administrative data. These data will be part of a Central Frame Data Base (CFDB) from which economic surveys will draw samples.

Administrative data will also be used to maintain the CFDB. This and other elements of the reorganization strategy are contained in Colledge and Lussier (1985). Experiences in the implementation of the strategy are contained in Colledge (1987).

One of the first steps was to formulate definitions of the CFDB units. A fundamental unit is the business entity. A business entity is defined in Statistics Canada (1987) as 'an economic transactor having the responsibility and authority to allocate resources in the production of goods and/or services, thereby directing and managing the receipt and disposition of income, the accumulation of property, the borrowing and lending of capital, and the maintenance of complete financial statements accounting for their responsibilities'.

The Central Frame Data Base currently being built by Statistics Canada attempts to represent the structure of the Canadian economy. It recognizes that this economy is dominated by a small number of large business entities who account for the majority of the activity within the economy. The CFDB is divided into two components paralleling this dichotomy.

One component, the Integrated Portion (IP), provides coverage of the small number of large or otherwise important business entities, while the other, the Non-Integrated Portion (NIP),

¹ Colleen Clark, Social Survey Methods Division, Statistics Canada, 4-C1 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6; and Robert Lussier, Business Survey Methods Division, Statistics Canada, 11-M R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

covers the remaining large number of smaller entities. The entities in the former component are more complex. Hence, the identification of those portions of the complex business entity that are of interest to a particular survey requires substantial effort.

The Integrated Portion (IP) of the CFDB attempts to represent the complex structure of business entities through the use of an Information Model. The model consists of five structures linked together which describe a business entity. These structures allow survey populations to be accurately identified. The five structures are:

- i. The *legal structure* which describes the legal representation of the business entity. It is comprised of legal entities and their relationships of ownership and control. Examples of legal entities are incorporations under federal or provincial charter.
- ii. The *operating structure* which describes how the business entity operates and how it organizes its accounting system. It is comprised of operating entities. This structure organizes and controls the production of goods and/or services. It is an attempt to structure the business entity as it sees itself. Examples of operating entities are divisions, profit centres, and plants.
- iii. The *statistical structure* which consists of a hierarchy of statistical entities. These entities are derived from the associated operating structure depending on the units within the operating structure for which records for a particular set of data are maintained.
- iv. The *reporting structure* which consists of reporting arrangements for each selected statistical entity by survey. The data available in the accounting system of the business entity are collected from the reporting entities.
- v. The *administrative structure* which contains administrative data such as income tax data collected from legal entities and payroll deduction account data collected from operating entities.

Entities on the statistical and reporting structures are generated by Statistics Canada for the purpose of collecting, editing, estimating, and tabulating economic data. The entities on the other three structures are externally defined.

The complex process of determining the boundaries of the business entity and of delineating its five IP structures and their associated links is termed 'profiling'. This network view of the business entity is the 'profile'. The data to construct a profile are obtained through a contact with the business entity or some component of it. The entity's legal and operating structures as well as some administrative structure data items are obtained, or, reviewed and updated during the interview. The statistical structure is then generated or updated automatically from the new operating structure. Finally, default reporting entities are created for new selected statistical entities using selected fields from the legal, operating or administrative structures. These entities may subsequently be updated as a result of the first survey contact with the respondents or of special arrangements negotiated with the respondents.

The type of profiling contact used depends on the entity's complexity and any special reporting arrangements. The most complex and important entities will receive a personal visit from either Head Office or Regional Office personnel. The remaining entities will be contacted by telephone. Entities will be contacted about once every two years, or more often, depending on how quickly their structures change.

Cyclical profiling, whereby business entities are periodically contacted, is one method that will be used to keep the IP of the CFDB current. A survey feedback process and data from administrative sources will also be used.

The design and construction of the CFDB is taking place over three years culminating in a data base that will be available for integration into survey programs. At implementation stage,

most of the data in the Integrated Portion of the CFDB should have come from a profiling process that began in April 1986. However, no single list of business entities in-scope for a profile was available in April 1986.

Administrative data played a major role in initiating the profiling process. It was used as a starting point to construct the current Statistics Canada view of the business entity. A list of business entities in-scope for an initial profile was assembled from administrative data sources. Section 2 describes how this was accomplished. Section 2.1 gives the frame requirements. A description of the data sources used to build the frame follows in Section 2.2. Section 2.3 shows how the frame unit was constructed and how the various data sources were combined to build the frame.

Section 3 describes how administrative data will be used to detect potential changes in a business entity and then to initiate the maintenance profiling process. The results of a simulation study done to quantify the proposed use of administrative data sources are then presented. The paper concludes with a discussion of several issues that this study has raised.

2. USE OF ADMINISTRATIVE DATA FOR INITIAL PROFILING

2.1 Frame Requirements

The first step in building the frame for initial profiling was to define the frame unit. The ideal one would be the business entity. However this entity was not available either internally or externally to Statistics Canada. The units available to us were essentially legal entities. It was necessary, then, to group legal entities to approximate business entities. The frame unit was defined as a grouping of legal entities subject to the following constraints:

- i. The definition of the business entity implies that it covers all legal entities linked through ownership where ownership is defined as the owning more than 50% of the voting rights of a legal entity. The grouping of legal entities through this ownership rule is restricted to one level of foreign ownership outside Canada.
- ii. There has to be a single Canadian legal entity that owns all other Canadian legal entities in the business entity. This is necessary because profiling contacts with the business entity could only be made in Canada.

The next step was to determine which frame units would comprise the frame and what data was required for each. The frame from which business entities would be selected for an initial profiling contact and from which the initial picture of the business entity would be generated would contain all business entities in-scope for a contact.

Business entities are in-scope for a profiling contact if they qualify to be members of the Integrated Portion of the CFDB. Membership is determined by criteria applied to the legal structure that describes the legal representation of the business entity.

Legal structures can become members of the Integrated Portion in one of two ways. First, if the structure consists of only one legal entity then the legal entity is part of the Integrated Portion if its revenue during its fiscal year of interest is above a prespecified value. This prespecified value depends on the legal entity's major industry and the location of its head office. Alternatively, if the legal structure consists of more than one legal entity then the legal structure is part of the Integrated Portion if at least one of the legal entities in the structure has a revenue above its appropriate prespecified value.

Therefore, in order to determine which business entities are in-scope, the following information was required for every legal entity:

- i. Relationships of ownership between legal entities.
- ii. Revenue in the fiscal year of interest, primary industry, and head office location.

For business entities that qualify to be on the frame and, hence, to receive an initial profiling contact, information was required to select and contact the entity. The following was required to select the entity:

- i. All industries in which the business entity was involved so that the Wholesale and/or Retail industries could be contacted first. The surveys of these industries required a set of statistical entities that had been generated from a profiling contact before other surveys did.
- ii. The number of physical locations of all business entities that consist of one legal entity or that consist of two legal entities of which the owner is foreign. This data item determined the type of profiling contact that would be made as either a telephone contact by Regional Office staff or a personal visit by Regional or Head Office staff.
- iii. The province in which the ultimate Canadian corporate ownership was based. The province was used to distribute the workload of making the profiling contacts to regional offices according to their capacities.

In order to contact the business entities, name and address were required for the legal entity at the top (excluding foreign owners) of the business entity. Contact data and any special reporting arrangements that surveys had recently used would be desirable.

2.2 Data Sources

The data sources which could be used were restricted, primarily, by the frame coverage requirements. This restriction eliminated sample lists and many industry specific lists such as survey frames. Only data sources that were lists of all legal entities potentially in-scope for a profiling contact that carried, at least, some of the required data items could be considered. The data sources that could at least be partially integrated by computer were:

- i. The *Inter-Corporate Ownership Database* (ICO) which is a list of all legal entities operating in Canada that are owned by either foreign or Canadian legal entities and their owners. The coverage of foreign legal entities is required to determine the ultimate owner.
- ii. The *Current Business Register* (BR) which is primarily a list of all legal entities that are employers. The number of physical locations of a legal entity, contact data (address and reporting arrangements) used by surveys, and the industries in which the legal entity operates are available here.
- iii. The *Corporation Tax Base* (CORP) which is a list of all legal entities that filed a corporate tax return with Revenue Canada, Taxation in a given year. The primary industry, the location of the Head Office, and revenue for the fiscal year are carried on this data source.
- iv. The *Individual Tax Base* (IND) which is a list of all individuals who filed a tax return with Revenue Canada, Taxation in a given year. Individuals who report self-employed income on their return are legal entities of interest to Statistics Canada economic surveys. Primary industry data and contact data are available from this tax base for each individual reporting self-employed revenue as is his/her revenue from self-employment.

Both of the tax base data sources (CORP and IND) are administrative data files. Administrative data received monthly from Revenue Canada, Taxation regarding an employer's payroll deductions are used to update the BR. The ICO data source is a census survey response file.

None of these data sources provides complete coverage and all the required data items. Rather, coverage could only be obtained by combining these data sources. The same is true for some required data items while for the rest more than one source could provide them. The strategy used to combine these data sources to obtain the best coverage and data quality is presented in the next section.

A fifth source, the *Quarterly Survey of Financial Statements* provided information on legal entities that prepare consolidated financial statements. This source was used in manually refining the business entities on the frame.

2.3 Frame Creation Procedures

The challenge in creating the frame for initial profiling contacts lay in integrating four data sources that had each been designed for different purposes and had never been integrated to this extent before. This situation is common to users of administrative data. The task was even more complex because this was the first time many concepts established for the CFDB were applied.

The constraints of limited time and resources forced the project team to make some assumptions when creating the frame. However, the assumptions were justifiable since the picture used on the frame would be corrected through the profiling process. A simple description of the procedures used is presented in this section.

There were three steps in the frame creation process, each of which is discussed in the following sections.

- i. Construct a list of all potential frame units;
- ii. Determine which are in-scope; and
- iii. Acquire selection and contact data.

2.3.1 Create Potential Frame Units

The frame unit was constructed by grouping legal entities in the following manner to create business entities. The legal entities were first grouped into legal structures. One legal structure consisted of that set of legal entities related via ownership of more than 50%. Relationships involving foreign legal entities were accepted only if the foreign legal entity owned or was owned by a Canadian legal entity. When a foreign entity owned more than one Canadian entity, the legal structure was divided into as many business entities as there were Canadian entities directly owned by the foreign entity. In this way, a profiling contact would be made with the ultimate Canadian owner of each resulting business entity. Examples are provided in Figure 1.

Individuals who reported self-employed income were considered as a legal structure containing only one legal entity. The ownership of corporations by individuals as well as relationships of joint venture between corporations were not considered in constructing business entities.

Therefore, we can think of the set of business entities in-scope for an initial profiling contact as two mutually exclusive groups. The first group consists of legal entities that represent individuals who report self-employed income. The Individual (IND) tax base contains a list of all potential frame units in this group.

The second group consists of legal entities that represent corporations operating in Canada. The Inter-Corporate Ownership (ICO) data source was manipulated to provide a list of corporations that belonged to legal structures containing more than one legal entity. A list of all legal entities that are not owned by any other legal entity was obtained from the Corporation tax base after elimination of those legal entities that were owned by other legal entities or were

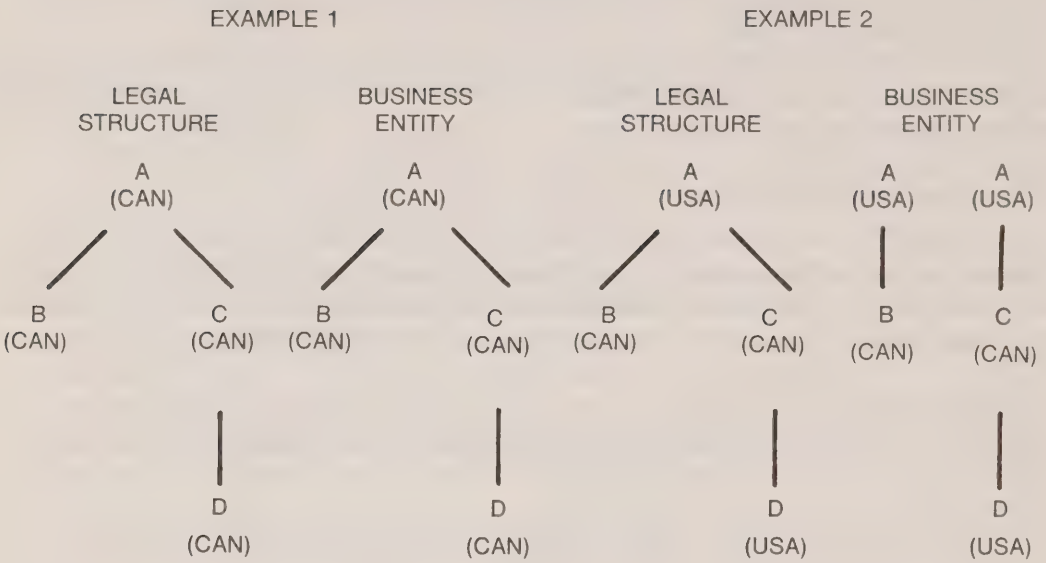


Figure 1. Defining Business Entities

owners themselves. That is, it was necessary to match the ICO source and the CORP Tax base to identify the overlap between them. Legal entities that appeared on both sources could thus be identified to ensure that they would only appear once on the frame. Linkage between the two sources was not straight- forward and involved a clerical process because a common identification number was often not available.

2.3.2 Determine In-Scope Frame Units

The data required to determine if individuals reporting self-employed income were in- scope was on the IND tax base. It was a simple step to determine if a legal entity was above its appropriate prespecified cut-off.

The situation was more complex for corporations. The linkage achieved between ICO and CORP provided the data required to apply the cut-off rule. However, about 20% of the corporations on ICO could not be linked to CORP. In these cases an assumption was made which led to an overestimation of the set of business entities in-scope for an initial profile. It was assumed that legal structures which contained at least one unlinked corporation satisfied the frame inclusion conditions. Otherwise, legal structures were frame members if at least one corporation satisfied the cut-off rule.

2.3.3 Acquire Selection and Contact Data

The result of the previous step was a proxy list of all business entities in-scope for an initial profiling contact. The data required for selection and contact described in Section 2.1 that are not already on the frame were available from the BR. The frame and the BR overlap because a majority of the frame units representing corporations and a smaller proportion of the frame units representing individuals are employers. Linkage between the frame and the BR was required so that data from the BR could be added to the frame for units found on both sources. That is, it was necessary to detect duplication between the two sources.

It was even more difficult to link these two sources than it had been to link the ICO and CORP sources. This was due not only to the frequent absence of common identification numbers as in the ICO-CORP case but also because the BR resembles a business entity's operating structure more than its legal structure. The name and address from the BR were used for linking when no common identification number was available. However, the names and addresses on the BR often refer to 'trade' or 'operating' locations which are sometimes different from the 'legal' names and addresses on the ICO and CORP sources. When this occurred it was difficult to establish a link and hence eliminate duplication.

There were some frame units for which no link to the BR was achieved either because they were non-employers and therefore not on the BR or the linkage procedures could not establish the link. In these cases subsequent stages in the initial profiling process were amended to accommodate the frame limitations. Contact data of a lesser quality were taken from the tax base. The selection criteria were changed to reflect the absence of data on industrial breakdown and physical locations for these legal entities.

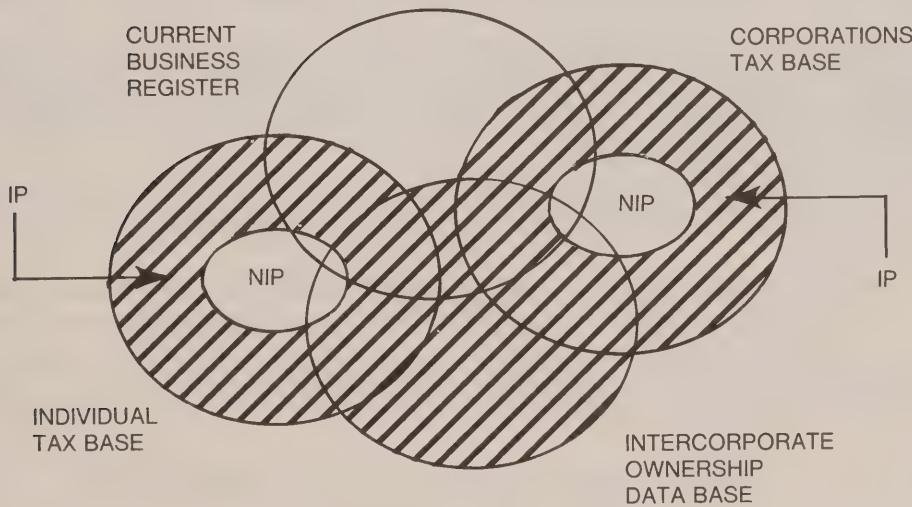


Figure 2. Frame for Initial Profiling

When a legal entity was involved in only one industry, the primary industry was available from both the tax bases and the BR. It was necessary, then, to reconcile this common data item when they were different. In this case the BR industry was used since it was considered more reliable.

A pictorial representation (not to scale) of the resulting frame is shown in Figure 2.

2.3.4 Evaluate the Frame

The quality of the resulting frame was assessed by three projects. First, the consistency of the frame with the specifications for creating it was verified.

The second project involved comparing various distributions of the legal entities on the frame with the same distributions produced from an independent simulation of the Integrated Portion. The distributions did not differ significantly.

Lastly the frame was assessed by comparing it with the BR. A sample of 30 of the larger units in the BR was matched to the frame for initial profiling. All of the entities were found but with great difficulty because the two sources use different concepts.

2.4 Conclusion

The frame strategy just described was based on some simplistic assumptions regarding coverage, data quality, and the way in which business entities operate. 'Shortcuts' were often used to satisfy the frame requirements. It was felt that this approach was justified because of the role of the frame as a provider of initial pictures of business entities that would be updated during the profiling process. The implications of making these assumptions are discussed in this section.

The population of business entities in-scope for an initial profiling contact may contain duplicates and out-of-scope units. If so, then more profiling contacts than necessary will be made. This would increase Statistics Canada's production costs. It would unduly burden the respondent with duplicate requests. Finally, the image of Statistics Canada could be adversely affected.

The population may be underestimated. Nevertheless, the missing units will be profiled at a later date. This would delay the introduction of new large units into the Integrated Portion of the CFDB. The missing units would be covered by the Non-Integrated Portion in the interim rather than the Integrated Portion.

Inaccurate selection and/or contact data could complicate or delay contact until accurate data could be found. The consequence in these cases is also an inaccurate CFDB until the profile is completed.

These experiences demonstrate the complications introduced when administrative data are used. They also illustrate the care that must be taken in ensuring the compatibility of administrative data with one's requirements. Examples were provided of the types of ensuing compromises that must be made when such a compatibility cannot reasonably be reached.

3. USE OF ADMINISTRATIVE DATA IN SUBSEQUENT MAINTENANCE PROFILES

3.1 Cyclical and Reaction Profiling

There will be two types of subsequent maintenance profiling, namely cyclical and reaction profiling. Each of these is explained below.

Cyclical profiling is the process that will ensure that all business entities in the profile population get reprofiled within a certain period of time. It is expected according to current budget forecasts that this period of time will be two years. Time elapsed since the business entity's last profile will be the factor that determines eligibility for cyclical profiling. Other factors will be taken into account to prioritize the eligible units within cyclical profiling.

Reaction profiling is the process that will profile a business entity as a result of information through a source other than profiling that changes may have occurred to that business entity and that the statistical image of the business entity on the register may not be valid any longer. Reaction profiling will keep the CFDB more up-to-date than if only the cyclical profiling mechanism were used. Some of the sources of information on changes are the various files of administrative data received regularly at Statistics Canada.

3.2 Sources of Administrative Data That Can be Used

The three sources of administrative data that Statistics Canada can use to update its central register that are discussed in this paper are:

- the Individual Tax Base;
- the Corporation Tax Base; and
- data on payroll deduction accounts captured by the tax authorities.

Generally, individuals and corporations file a single tax return for a reference year. However, it is possible to have more than one return for a reference year if, for example, a corporation changed its fiscal year end with the approval of the tax authorities. Nevertheless, one can say that tax returns are an annual source of changes.

The receipt of the tax bases at Statistics Canada does not occur at a single point in time. In fact, Statistics Canada receives files of tax data regularly for a reference year over a period of two years. Thus, one could perform monthly updates to the register from tax data but each register record would generally be updated only once a year.

On the other hand, an employer is generally expected to send remittances for his payroll deduction accounts on a monthly basis. In turn, Statistics Canada receives a file of payroll deduction account data once a month. Thus, monthly updates can be made to the register from payroll deduction account data and each register record can in theory be modified every month.

Note that there are other sources of administrative data that could be used. They are not discussed in this paper because they are not obtained on a universe basis or on a regular basis. They are nevertheless worth mentioning. These are:

- limited information on corporations that have not filed a tax return but are believed to be active, captured by the tax authorities;
- additional data captured from a sample of tax returns by Statistics Canada; and
- data on a tax authority form filled out by employers when they request a payroll deduction account, captured by Statistics Canada.

3.3 Signals of Change

Signals of change were developed from the administrative sources described in the previous section. These signals identify administrative records for which changes to their associated statistical entities may have occurred. They also inform the register that reaction profiling may be desirable for these entities to keep the register up-to-date.

Table 1
Signals by Administrative Source

Administrative Source	Number Of Distinct Signals	Examples
Annual Individual Tax Returns	50	Change from single province of taxation to multiple jurisdiction
Annual Corporation Tax Returns	49	Start of a joint venture
Monthly Payroll Deduction Accounts	38	New account with descriptions in the name that identify a corporation

The signals are administrative source dependent. For each of the three sources listed in 3.2 the signals consist of comparison tests between new data received for an administrative record and the last data received for the same record from the same source. These tests may involve a single field or a group of fields and may be conditional on a single field or several fields. These comparison tests attempt to identify real world events that have an impact on the statistical entities and not only on the administrative entities. Remember that the statistical entities exist for the purpose of economic statistical programs and often are completely different from the legal-administrative reality. Therefore, these comparison tests should optimize the detection of changes in the administrative data that reflect a change in the statistical entities. As an example, change of ownership of a manufacturing plant may mean the death of an administrative record and the birthing of a new one. On the statistical entities, it may however mean no change as the same establishment with its capabilities to provide the required data may still exist.

If the frame was updated directly from the changes noted in the administrative records, the consequence would be a high incidence of apparent deaths and births in the statistical entities and a risk of incomplete or duplicated coverage. Thus there is a requirement to contact respondents, or at least to perform in-house research using all available documentation, to find out for signaled administrative records what happened to the statistical entities. The “translation” process is not trivial at all and its resolution constitutes the purpose of reaction profiling.

The number of signals that were determined from each source together with some signal examples are presented in Table 1. One should however note the following points in studying the data on the number of signals. Some signals are very refined while others are not. It was often decided to split an original signal into mutually exclusive sub-signals because it was felt that it may be more informative in determining the action to take from the signal. The most trivial example concerns the Payroll Deduction Accounts. Eighteen of the 40 signals represent changes in the estimated number of employees covered by the account. The 18 signals distinguish between increases and decreases in the estimated number and the magnitude for each of them. It was thought that such a breakdown would be informative to prioritize the clerical work. Nevertheless one could consider these signals as one.

It is expected that even though tax returns are processed regularly, a given return will generally generate signals at most once per reference year while a given payroll deduction account may generate a signal or signals every month. What is of more interest therefore is not the number of signals defined per source but the number of records that are identified by these signals. This would give an idea of the amount of clerical resources that will have to be invested to update the register from administrative sources. A simulation study was thus undertaken to address this issue.

3.4 Simulation Study

The simulation study consisted of applying the signals previously described to the following populations:

- the individual tax returns for fiscal periods that ended in 1984 to detect changes that had taken place during these periods;
- the corporation tax returns for fiscal periods that ended in 1984 to detect changes that happened during these periods;
- the payroll deduction account of the beginning of October 1985 to detect changes that had occurred since the beginning of September 1985.

The results of the simulation study are presented in Table 2. The following observations can be made on the results:

- There are a very large number of tax returns that generate signals: only about one eighth of the individual tax returns and one fifth of the corporation tax returns do not generate any signals.
- There are 8,258 payroll deduction accounts that generated signals for a one month period. If one supposes uniformity of the payroll deduction account signals over months, there would be almost 100,000 accounts signaled in a year. Note that it is likely that accounts would be signaled in more than one month and therefore there would be duplicates if one cumulated the signals.
- If all records signaled in a year are added, it gives the grand total of 244,269 signaled records. However, it is obvious that signals are duplicated between the administrative sources. For example, a change to the legal name of a business could be found on the tax return as well as on each of its payroll deduction accounts.

3.5 Questions Raised

The results of the simulation study as well as an examination of the role of the signals raise a certain number of issues with respect to the profiling activities.

Six of these issues are presented bpose.

Table 2
Results of Simulation Study

Administrative Source	Number In The Profile Population	Number Signaled	Percentage Signaled
Individual Tax Returns	72,190	63,446	87.9
Corporation Tax Returns	102,688	81,727	79.6
Payroll Deduction Accounts	134,973	8,258	6.1

3.5.1 Performance of Signals in Detecting Change(s) to Statistical Entities

The signals will attempt to flag legal and/or operating entities involved in real world events that have an impact on the statistical entities. An update will then be necessary on the central register to maintain the quality of the statistical products. Are the signals really reflecting real world events that affect the statistical entities or are there some that have no impact? If some are useless, work will be generated for no purpose.

A small-scale survey was conducted in 1986 to determine the usefulness of the signals with respect to the detection of changes to the statistical entities. However, for various reasons, the only signals that could be used were those of the simulation study. They refer to changes between tax returns of taxation years 1983 and 1984. Thus the time lag between the reference period of the signals and the survey period (1986) gave recalling difficulties to the respondents. This led to the inclusion of events which took place after the period as well as the omission of events which did occur in the reference period. The survey was therefore inconclusive and no other attempt has been made since then.

3.5.2 Repetitiveness of Signals

Signals will be received over time and from different independent sources. The tax returns in particular suffer from noticeable time delays. As a given signal is received, the CFDB may have already been updated to reflect the real world event behind the signal. This update may have been the result of processing a signaled record from another source or of conducting cyclical profiling or of incorporating feedback received from surveys. Therefore, signals cannot be treated independently of the CFDB to decide to perform a reaction profile. However, how should a signal be checked against the CFDB to see if the CFDB was already updated? As an example, if a large increase in revenue is flagged on a corporation tax return, how should one check if the CFDB was already updated to reflect the real world event behind this increase when one does not know the real world event behind it?

3.5.3 Omission of Signals of Change

Similarly, some records will not get signaled. Will the absence of signals definitively mean that no real world event occurred that need the statistical structure to be updated? Should other signals be developed to cover omissions? Again, the survey previously mentioned was inconclusive in answering these questions.

3.5.4 Availability of Resources to Handle Signaled Records

As the simulation study showed, a large number of records will be signaled. These will require manual work. It is likely, that there will not be sufficient resources to perform all this work. How should the total amount of resources to be devoted to reaction profiles be determined and how should this total amount be used to handle the signaled records? If constraints on resources demand that some signals be ignored, how will these be determined?

3.5.5 Response Burden

The results of the simulation study suggest that businesses will be contacted more often than every second year to check for frame changes other than through regular survey activity. This will increase response burden. Can a trade-off be established between increase in response burden and out-datedness of the register? What should this trade-off be?

3.5.6 Role of Cyclical Profiling

The large amount of records signaled by the tax returns in the simulation study raises a question about the usefulness of cyclical profiling. The number of records subject to cyclical profiling and not to reaction profiling can be deduced to be very small. First, suppose the results of the simulation study in terms of numbers hold for a second year. Then suppose the records signaled in the second year are not all the same in the first year but that there are new records signaled and that there are last year's records not signaled the second year. Then it can be safely assumed that the number of records which will not get a signal over two years will be very small. There may be only a few records left which will not be signaled on either one or the other year. This will in fact represent the maximum target population for cyclical profiling. Will it be necessary to perform a profile for these entities, knowing that they are not signaled by the Payroll Deduction Accounts nor by the tax returns?

4. CONCLUSION

Section 2 has shown how administrative data were used to build a frame for initial profiling. Administrative data offered extensive coverage. However, it was also seen that conceptual differences between one's requirements and administrative data can lead to complications requiring simplifying assumptions and compromises.

The resulting frame supported the initial profiling of all business entities except the most complex ones. In these cases the approximation given by the frame could not be accepted. Rather, extensive research was conducted on each business entity using elements such as public annual reports and survey responses.

The frame also played an important role in initializing the CFDB. It was used along with the Business Register to identify the members of the Integrated Portion.

The method by which administrative data will be used to initiate a maintenance profile was described in Section 3. Signals of change will be derived from various administrative sources and will generate requests to verify profiles. Many issues were raised in this respect. These issues are being addressed by the various design teams responsible for implementing the CFDB update strategy. A solution being investigated to solve some issues is to prioritize signals depending for example on the length of time since the entity was last profiled. Another solution is to develop a self-learning process. Experience will dictate which signals are useful and should be kept. Therefore, substantial work is still required before the process stabilizes in production.

REFERENCES

- COLLEDGE, M., and LUSSIER, R. (1985). A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 123-131.
- COLLEDGE, M. (1987). The Business Survey Redesign Project — Implementation of a New Strategy at Statistics Canada. Presented at the U.S. Bureau of the Census Third Annual Research Conference.
- STATISTICS CANADA (1987). Version 4.2 of the CFDB Data Dictionary. Business Survey Redesign Project Working Paper, March 4, 1987.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation
- 1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.3

Les remerciements doivent paraître à la fin du texte.
- 1.4

Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé
- Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction
- 3.1

Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5

Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, l).
- 3.6

Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux
- 4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie
- 5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

BIBLIOGRAPHIE

- COLLEDGE, M., et LUSSIER, R. (1985). A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 123-131.
- COLLEDGE, M. (1987). Projet de remaniement des enquêtes-entreprises – Mise en pratique d'une nouvelle stratégie à Statistique Canada. Document présenté à la *Bureau of the Census Third Annual Research Conference*.
- STATISTIQUE CANADA (1987). Version 4.2 of the CFDB Data Dictionary. Projet de remaniement des enquêtes-entreprises, document de travail, 4 mars 1987.

3.5.5 Fardeau de réponse

Les résultats de l'étude de simulation semblent indiquer qu'il faudra contacter les entreprises plus souvent qu'une fois tous les deux ans si l'on veut vérifier les changements à apporter à la base de sondage par d'autres moyens que les enquêtes régulières. Cela augmentera nécessairement le fardeau de réponse. Peut-on trouver un compromis entre l'obligation d'augmenter le fardeau de réponse pour tenir le registre à jour et la nécessité d'empêcher que le registre ne cesse d'être à jour? Quel devrait être ce compromis?

3.5.6 Rôle du processus d'établissement cyclique des profils

Le grand nombre d'enregistrements signalés par les déclarations d'impôt dans l'étude de simulation amène à s'interroger sur l'utilité du processus d'établissement cyclique des profils. On peut déduire des résultats de cette étude que très peu d'enregistrements devront être soumis au processus cyclique plutôt qu'au processus ponctuel. Supposons par exemple que les résultats de l'étude de simulation concernant le nombre d'enregistrements signalés la deuxième année vrait une deuxième année. Supposons ensuite que les enregistrements signalés la première année ne sont pas tous les mêmes que les enregistrements signalés la deuxième année, mais qu'un certain nombre de nouveaux enregistrements sont signalés et que certains enregistrements de la dernière année ne sont pas signalés la deuxième année. On peut alors supposer sans grand risque d'erreur que très peu d'enregistrements ne recevront aucun signal pendant deux ans. Il restera très peu d'enregistrements qui n'auront pas été signalés l'une ou l'autre année. Ce nombre représentera en fait le nombre maximum d'unités de la population à soumettre au processus d'établissement cyclique des profils. Faudra-t-il établir absolument le profil de ces entités, sachant qu'elles ne sont signalées ni par les comptes de retenue sur la paye, ni par les déclarations d'impôt?

4. CONCLUSION

Dans la partie 2, nous avons montré comment les données administratives ont été utilisées pour construire une base de sondage en vue de l'établissement de profils initiaux. Les données administratives offraient un taux de couverture très élevé. Toutefois, nous avons aussi vu que les différences conceptuelles existant entre nos besoins et les données administratives amenaient des complications qui nous ont obligés à faire des hypothèses et des compromis simplificateurs.

La base de sondage que nous avons créée incluait les éléments nécessaires à l'établissement des profils initiaux de toutes les entités commerciales à l'exception des plus complexes. Dans ces derniers cas, on ne pouvait pas accepter les approximations données par la base de sondage. Il a donc fallu faire beaucoup de recherches sur ces entités commerciales en utilisant d'autres sources d'information comme des rapports annuels publics et des réponses à des enquêtes.

La base de sondage a également joué un rôle important dans la construction et la mise en vigueur de la BDRC. Elle a servi avec le Registre des entreprises à déterminer quelles entités seraient incluses dans la partie intégrée.

La méthode suivant laquelle les données administratives seront utilisées pour établir les profils de mise à jour a été décrite dans la partie 3. Des signaux de changement seront tirés de diverses sources de données administratives et produiront des demandes de vérification des profils. Beaucoup de questions ont été soulevées à cet égard. Les diverses équipes chargées de mettre en pratique la stratégie de mise à jour de la BDRC s'emploient actuellement à trouver des solutions à ces questions. Une des solutions à une partie des problèmes serait de classer les signaux par ordre de priorité selon, par exemple, la durée de la période depuis la dernière fois qu'on a établi le profil d'une entité. Une autre solution serait de mettre au point un système autodidacte. L'expérience dira quels signaux sont utiles et devront être conservés. Il reste donc encore beaucoup de travail à faire avant que le processus soit tout à fait opérationnel.

3.5 Questions soulevées

Les résultats de l'étude de simulation et l'analyse du rôle des signaux soulèvent un certain nombre de questions au sujet de l'établissement des profils.

Six de ces questions sont présentées ci-dessous.

3.5.1 Dans quelle mesure les signaux permettent de détecter les changements survenus dans

les entités statistiques

Les signaux indiqueront plus ou moins fidèlement quelles entités juridiques et/ou exploitantes ont subi des changements réels qui ont un effet sur les entités statistiques. Il faudra ensuite mettre à jour le registre central pour maintenir la qualité des produits statistiques. Les signaux reflètent-ils réellement les événements du monde réel qui influent sur les entités statistiques ou est-ce qu'il y en a qui ne correspondent à aucun effet? Si certains n'ont pas de signification, cela va engendrer du travail inutile.

Une enquête à petite échelle a été menée en 1986 pour déterminer dans quelle mesure les signaux permettaient de déceler les changements survenus dans les entités statistiques. Toutefois, pour diverses raisons, les seuls signaux qu'on a pu utiliser étaient ceux de l'étude de simulation. Ces signaux ont trait aux changements survenus dans les déclarations d'impôt entre les années d'imposition 1983 et 1984. Mais le décalage entre la période de référence des signaux et la période d'enquête (1986) a posé des problèmes de mémoire aux répondants. C'est ainsi que les répondants ont inclus des événements qui ont eu lieu après la période de référence ou ont oublié des événements qui ont eu lieu pendant la période de référence. On a donc jugé que l'enquête n'était pas concluante, et aucun autre essai n'a été tenté depuis ce temps-là.

3.5.2 Répétition des signaux

Les signaux seront reçus à plusieurs moments et de sources différentes indépendantes les unes des autres. Les déclarations d'impôt sur le revenu, en particulier, entraînent des délais sensibles. Au moment où un signal est reçu, la BDRC pourra déjà avoir été mise à jour pour tenir compte de l'événement du monde réel indiqué par ce signal. Cette mise à jour pourra avoir été faite à l'occasion du traitement d'un signal venu d'une autre source ou du processus cyclique d'établissement des profils ou encore de l'incorporation de renseignements tirés d'autres enquêtes. Par conséquent, on ne peut pas traiter les signaux indépendamment de la BDRC pour décider de procéder à l'établissement ponctuel d'un profil. Toutefois, comment devrait-on vérifier un signal en se référant à la BDRC pour voir si la BDRC a déjà été mise à jour? Par exemple, dans le cas où une déclaration d'impôt sur le revenu des sociétés indique une grosse augmentation de revenu, comment peut-on vérifier que la BDRC a déjà été mise à jour pour tenir compte de l'événement du monde réel sous-jacent à cette augmentation quand on ne sait même pas quel est cet événement?

3.5.3 Omission des signaux de changement

De même, certains enregistrements ne seront pas signaux. Est-ce que l'absence de signaux signifiera nécessairement qu'aucun événement du monde réel ne s'est produit qui oblige à mettre à jour la structure statistique? Devrait-on mettre au point d'autres signaux pour couvrir les omissions? Encore ici, l'étude mentionnée précédemment n'a pas donné de réponses concluantes à ces questions.

3.5.4 Possibilité d'obtenir des ressources en quantité suffisante pour traiter les enregistrements signaux

Comme l'étude de simulation l'a montré, beaucoup d'enregistrements seront signaux qu'il faudra traiter en partie à la main. Il est probable qu'il n'y aura pas assez de ressources pour accomplir tout ce travail. Comment devrait-on procéder pour déterminer quelle proportion de l'ensemble des ressources devrait être consacrée à l'établissement ponctuel de profils et pour déterminer la façon d'utiliser les ressources pour traiter les enregistrements signaux? Si une rareté de ressources oblige à ignorer un certain nombre de signaux, comment allons-nous déterminer quels signaux il faut laisser tomber?

Même si les déclarations d'impôt sur le revenu sont dépouillées régulièrement, on s'attend qu'une déclaration donnée produise en général des signaux au maximum une fois par année de référence, tandis qu'un compte de retenue sur la paye pourrait produire un ou plus d'un signal tous les mois. Mais ce qui nous intéresse le plus, ce n'est pas le nombre de signaux que peut transmettre une source, mais le nombre d'enregistrements que ces signaux permettent de trouver. Cela donnerait une idée des ressources en personnel de bureau à investir pour mettre à jour le registre à partir de sources administratives. On a donc fait une étude de simulation pour répondre à cette question.

3.4 Étude de simulation

L'étude de simulation a consisté à appliquer les signaux décrits précédemment aux populations suivantes:

- les déclarations d'impôt sur le revenu des particuliers pour les exercices financiers précédant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les déclarations d'impôt sur le revenu des sociétés pour les exercices financiers précédant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les comptes de retenue sur la paye du début d'octobre 1985, pour détecter les changements survenus depuis le début de septembre 1985.

Les résultats de l'étude de simulation sont présentés dans le tableau 2. On peut faire les observations suivantes sur les résultats obtenus.

- Beaucoup de déclarations d'impôt sur le revenu produisent des signaux: seulement un huitième environ des déclarations d'impôt sur le revenu des particuliers et un cinquième des déclarations d'impôt sur le revenu des sociétés n'ont pas produit de signal.
- Durant la période d'un mois observée, 8,258 comptes de retenue sur la paye ont produit un signal pour le mois considéré. Si l'on suppose que les signaux émis par les comptes de retenue sur la paye sont distribués uniformément d'un mois à l'autre, cela ferait presque 100,000 comptes signalés par année. Il convient toutefois de remarquer qu'il est fort possible qu'un même compte soit signalé plus d'un mois et qu'on risque donc de compter au moins deux fois les signaux si on les additionne sur une période d'un an.
- Si l'on additionne tous les enregistrements signalés au cours de l'année, cela donne au total 244,269 enregistrements signalés. Il est toutefois évident qu'un même signal peut se trouver dans plus d'une source administrative. Par exemple, un changement de raison sociale d'une entreprise peut être indiqué dans la déclaration d'impôt produite par l'entreprise et dans chacun de ces deux comptes de retenue sur la paye.

Tableau 2

Résultats de l'étude de simulation

Source administrative	Population totale	Nombre d'enregistrements signalés	Nombre d'enregistrements signalés en pourcentage de la population totale
Déclarations d'impôt sur le revenu des particuliers	72,190	63,446	87.9
Déclarations d'impôt sur le revenu des sociétés	102,688	81,727	79.6
Comptes de retenue sur la paye	134,973	8,258	6.1

Les signaux dépendent de la source administrative. Dans chacune des trois sources énumérées dans la partie 3.2, les signaux sont des tests de comparaison entre les nouvelles données reçues pour un enregistrement donné et les données précédentes reçues de la même source pour l'enregistrement en question. Les tests peuvent porter sur une seule zone ou sur un groupe de zones et être conditionnel à une ou plus d'une zone. Ces tests de comparaison tentent de refléter les événements du monde réel qui influent sur les entités statistiques et non pas seulement ceux qui influent sur les entités administratives. Il faut se rappeler que les entités statistiques ont été créées aux fins des programmes de statistiques économiques et souvent diffèrent complètement des entités juridiques ou administratives. Aussi, ces tests de comparaison devraient permettre le plus possible de trouver les changements survenus dans les données administratives qui reflètent un changement dans les entités statistiques. Par exemple, un changement de propriétaire d'une usine de fabrication peut signifier qu'un enregistrement administratif disparaîtra et qu'un autre apparaîtra. Cela peut par contre n'avoir aucun effet sur les entités statistiques étant donné que le même établissement avec le même degré de capacité de fournir les données requises peut continuer d'exister.

Si la base de sondage était mise à jour directement à partir des changements observés dans les enregistrements administratifs sans aucun autre contrôle, il y aurait une forte proportion de créations et de disparitions apparentes d'entités et un risque proportionnel de couverture incomplète ou de couverture en double. C'est pourquoi il faut aussi contacter les répondants ou au moins faire des recherches internes à l'aide de tous les documents qu'on peut obtenir pour trouver ce qui est arrivé aux entités statistiques dans le cas des enregistrements administratifs signalés. Le processus de «traduction» n'est pas ce qu'il y a de plus simple et la recherche d'une solution constitue justement l'objectif du processus d'établissement ponc-tuel de profils.

Le nombre de signaux retenus pour chaque source et quelques exemples de signaux sont présentés dans le tableau 1. Il faut toutefois noter les points suivants quand on regarde les données sur le nombre de signaux. Certains signaux sont très raffinés, d'autres pas. Dans bien des cas, on a décidé de diviser le signal initial en deux sous-sig-naux mutuellement exclusifs parce qu'on estimait que cela pouvait contribuer davantage à déterminer les bonnes mesures à prendre à partir du signal. L'exemple le plus simple concerne les comptes de retenue sur la paye. Dix-huit des quarante signaux représentent des changements dans le nombre estimatif d'employés couverts par le compte. Les 18 signaux permettent de distinguer entre des augmentations et des diminutions du nombre estimatif et tiennent compte de l'ampleur des augmentations et des diminutions. On a estimé que cette façon de procéder aiderait à classer par ordre de priorité les travaux manuels à accomplir. On pourrait néanmoins considérer tous ces signaux comme un seul signal.

Tableau 1

Signaux par source administrative

Source administrative	Nombre de signaux distincts	Exemples
Déclarations annuelles d'impôt sur le revenu des particuliers	50	Passage d'une seule à plus d'une province d'imposition
Déclarations annuelles d'impôt sur le revenu des sociétés	49	Début d'une coentreprise
Comptes mensuels de retenue sur la paye	38	Nouveau compte avec la description du nom qui identifie une société

L'établissement ponctuel de profils est le processus d'établissement de profils des entités commerciales qui est déclenché parce qu'on a reçu d'une autre source des renseignements selon lesquels des changements se seraient produits dans l'entité commerciale et que l'image statistique ponctuel de profils fera que la BDRC sera plus à jour que si l'on utilisait seulement le mécanisme d'établissement cyclique des profils. Les divers fichiers de données administratives reçus régulièrement à Statistique Canada font partie des autres sources d'information sur les changements.

3.2 Sources de données administratives pouvant être utilisées

Les trois sources de données administratives que Statistique Canada peut utiliser pour mettre à jour son registre central et dont nous parlons dans le présent document sont :

- la Base des unités de l'impôt sur le revenu des particuliers,
- la Base des unités de l'impôt sur le revenu des sociétés, et

- les données sur les comptes de retenue sur la paye saisies par les administrations fiscales.

En général, les particuliers et les sociétés produisent une seule déclaration d'impôt pour une année de référence donnée. Il se peut toutefois qu'ils en produisent plus d'une parce que, par exemple, une société aurait changé la fin de son exercice financier avec l'approbation des administrations fiscales. On peut quand même dire que les déclarations d'impôt sur le revenu sont une source de données annuelles sur les changements.

Ce n'est pas seulement une fois par année que les bases des unités de l'impôt sur le revenu sont communiquées à Statistique Canada. En fait, Statistique Canada reçoit régulièrement pendant deux ans les fichiers de données fiscales se rapportant à une année de référence donnée. Par conséquent, on pourrait mettre à jour tous les mois le registre à partir des données fiscales, mais chaque enregistrement du registre peut en théorie être modifié tous les mois.

Il convient de noter que d'autres sources de données administratives peuvent aussi être utilisées. On n'en parle pas dans le présent document parce qu'elles ne couvrent pas toute la population, ni sur une base régulière. Il convient toutefois de les citer. Ce sont :

- l'information limitée recueillie par les administrations fiscales sur les sociétés qui n'ont pas produit de déclaration d'impôt mais dont on pense qu'elles sont actives,
- d'autres données recueillies d'un échantillon de déclarations d'impôt sur le revenu par Statistique Canada, et
- les données recueillies par Statistique Canada à même les formules fiscales remplies par les employeurs demandant qu'on leur ouvre un compte de retenue sur la paye.

3.3 Signaux de changement

On a mis au point un système de signaux de changement qui proviendrait de chacune des sources de données administratives décrites dans la partie précédente. Ces signaux, rapportés à des enregistrements administratifs, indiquent quelles entités statistiques ont pu subir des changements. Ils indiquent également aux responsables du registre qu'il peut être souhaitable de refaire le profil de ces entités statistiques à l'aide du processus d'établissement ponctuel de profils pour tenir le registre à jour.

2.3.4 Évaluation de la base de sondage

La qualité de la base de sondage obtenue a été évaluée de trois façons. On a d'abord vérifié que la base de sondage concordait avec les spécifications développées pour sa création. Deuxièmement, on a comparé diverses distributions des entités juridiques figurant dans la base de sondage avec les mêmes distributions produites de façon indépendante à partir d'une simulation de la partie intégrée. Les distributions ne différaient pas de façon significative. Enfin, on a évalué la base de sondage en la comparant au RE. Un échantillon de 30 des plus grandes unités du RE a été apparié à la base de sondage créée en vue de l'établissement des profils initiaux. On a retrouvé toutes les unités, quoique très difficilement, parce que les deux sources n'utilisent pas les mêmes concepts.

2.4 Conclusion

La stratégie élaborée pour créer la base de sondage qu'on vient de décrire reposait sur des hypothèses simples au sujet de la couverture, de la qualité des données et de la façon dont fonctionnent les entités commerciales. On a souvent utilisé des « raccourcis » pour remplir les conditions requises par la base de sondage. On a estimé que cette approche était parfaitement justifiée étant donné qu'on jugeait que le rôle de la base de sondage était de donner des entités commerciales une première image qui pouvait ensuite être mise à jour pendant le processus d'établissement des profils. Les conséquences de ces hypothèses sont analysées dans les paragraphes qui suivent.

La population des entités commerciales pouvant faire partie du champ des entités susceptibles d'être contactées en vue de l'établissement d'un profil initial peut contenir des unités comptées deux fois et des unités hors du champ. Si c'est le cas, on prendra plus de contacts qu'il n'est nécessaire. Cela augmentera les coûts de production de Statistique Canada et accroîtra indûment le fardeau de réponse de certains répondants en les obligeant à répondre deux fois aux mêmes questions. Enfin, l'image de Statistique Canada pourrait être ternie un peu. La population pourrait être sous-estimée. Dans ce cas-là, on pourra établir le profil des unités manquantes un peu plus tard. Cela risque de retarder l'introduction de nouvelles grandes unités dans la partie intégrée de la BDRC. Entre-temps, les unités manquantes seraient couvertes par la partie non intégrée et non par la partie intégrée.

L'imprécision des données de sélection et/ou de contact pourrait compliquer ou retarder les contacts jusqu'à ce qu'on obtienne des données précises. Dans ces cas-là, il en résulte également que la BDRC sera imprécise jusqu'à ce que tous les profils aient été achevés.

Ces divers résultats montrent les complications que peut entraîner l'utilisation de données administratives. Ils montrent également qu'il faut bien s'assurer que les données administratives utilisées concordent avec les objectifs fixés. Des exemples ont été donnés des types de compromis qu'il faut faire quand on ne peut raisonnablement pas obtenir une telle compatibilité.

3. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DES PROFILS ULTÉRIEURS

3.1 Établissement cyclique et établissement ponctuel de profils

Il y aura deux types de mise à jour de profils initiaux, à savoir l'établissement cyclique de profils et l'établissement ponctuel de profils. Chaque type est expliqué ci-dessous.

L'établissement cyclique de profils est le processus par lequel on s'assure que le profil de toutes les entités commerciales de la population est refait au bout d'un certain temps. Compte tenu des prévisions budgétaires actuelles, on prévoit que cette période sera de deux ans. Le temps écoulé depuis le dernier profil sera le premier critère utilisé pour déterminer si cette entité doit être soumise au processus d'établissement cyclique des profils. On tiendra compte d'autres facteurs pour classer par ordre de priorité les entités dont il faut refaire le profil.

Il a été encore plus difficile d'appartier ces deux sources de données que les données de la base de données L.P.F et les données de la base IRS. Cela était attribuable non seulement au fait qu'on ne retrouvait souvent pas de numéros d'identification communs dans les deux sources, comme dans le cas L.P.F-IRS, mais aussi au fait que le RE ressemble plus à la structure opérationnelle des entités commerciales qu'à leur structure juridique. Le nom et l'adresse tirés du RE ont été utilisés pour appartier les enregistrements quand il n'y avait pas de numéro d'identification commun. Toutefois, les noms et les adresses figurant dans le RE s'appliquent souvent aux locaux «commerciaux» ou «exploitants» qui ont parfois des noms et des adresses différents des noms et des adresses «juridiques» figurant dans la base de données L.P.F et la base des unités IRS. Quand cela se produit, il est difficile d'établir un lien et donc d'éliminer les doubles comptes.

La base de sondage contenait certaines unités pour lesquelles on n'a pas établi de lien avec le RE soit parce que ces unités n'étaient pas des employeurs, et donc ne figuraient pas dans le RE, soit parce que les procédures de couplage n'ont pas permis d'établir le lien. Pour ces cas-là, les étapes ultérieures du processus d'établissement des profils initiaux ont été modifiées de manière à tenir compte des conditions imposées par la base de sondage. Des données de contact moins bonnes ont été tirées de la base des unités de l'impôt. Les critères de sélection ont également été modifiés de manière à tenir compte du fait qu'il puisse ne pas y avoir de données sur la répartition industrielle et les locaux d'affaires de ces entités juridiques. Quand une entité juridique était active dans plus d'une industrie, l'activité principale pouvait être déterminée à partir des bases des unités de l'impôt sur le revenu et du RE. Il a fallu alors concilier les éléments d'information communs quand ils différaient. Dans ce cas, l'industrie indiquée dans le RE a été utilisée étant donné que l'on jugeait cette source de renseignements plus fiable.

La figure 2 illustre par une figure (qui n'est pas à l'échelle) la base de sondage ainsi créée.

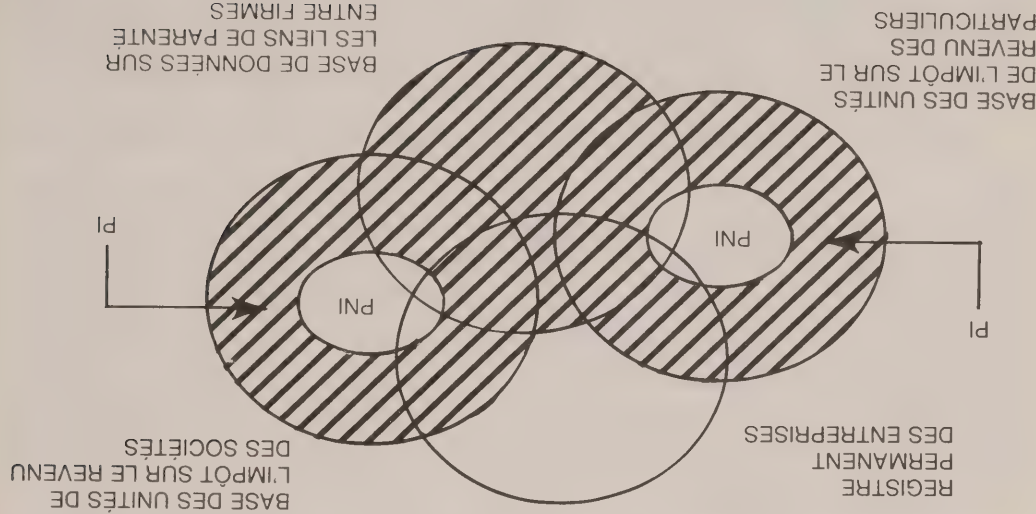


Figure 2. Base de sondage en vue de l'établissement des profils initiaux

Les particuliers ayant déclaré un revenu provenant d'un travail autonome étaient considérés comme une structure juridique contenant une seule entité juridique. Pour la construction des entités commerciales, on n'a pas tenu compte du fait que des sociétés puissent appartenir à des particuliers, ni des relations de coentreprise entre des sociétés en coparticipation.

Par conséquent, on peut considérer que l'ensemble des entités commerciales faisant partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial est formé de deux sous-ensembles mutuellement exclusifs. Le premier groupe est formé des entités juridiques qui repésentent des particuliers ayant déclaré un revenu provenant d'un travail autonome. La base des unités de l'impôt sur le revenu des particuliers (IRP) contient une liste de toutes les unités de ce groupe pouvant éventuellement faire partie de la base de sondage. Le deuxième groupe se compose des entités juridiques qui représentent les sociétés ayant des activités au Canada. La source de données sur les liens de parenté entre firmes (LPF) a été traitée de manière à fournir une liste des sociétés appartenant aux structures juridiques contenant plus d'une entité juridique. On a dressé une liste de toutes les entités juridiques appartenant pas à une autre entité juridique en éliminant de la base des unités de l'impôt sur le revenu des sociétés les entités juridiques qui appartenaient à une autre entité juridique ou qui étaient elles-mêmes propriétaires d'autres entités. Pour cela, il a fallu comparer les données de la base de données LPF aux données de la base IRS pour déterminer dans quelle mesure les deux bases se recoupaient. De cette façon, on a pu déterminer quelles entités juridiques figuraient dans les deux bases pour faire en sorte qu'elles n'apparaissent qu'une fois dans la base de sondage. Le couplage des entités des deux sources n'a pas été facile à effectuer et comportait une partie manuelle parce que souvent il n'existait pas de numéro d'identification commun aux enregistrements des deux sources.

2.3.2 Détermination des unités susceptibles de faire partie de la base de sondage

Les données dont on avait besoin pour déterminer si les particuliers ayant déclaré un revenu provenant d'un travail autonome pouvaient faire partie de la base de sondage figuraient dans la base IRP. Il a été très simple de déterminer si le revenu d'une entité juridique était supérieur ou non au seuil fixé au préalable pour cette entité.

La situation était plus compliquée pour les sociétés. Le couplage des données de la base de données LPF et des données de la base IRS nous a fourni les données dont nous avions besoin pour appliquer la règle du seuil limite. Toutefois, environ 20% des sociétés figurant dans la base de données LPF n'ont pu être liées à des sociétés de la base IRS. Dans ce cas, il a fallu faire une hypothèse qui a entraîné une surestimation du nombre d'entités commerciales pouvant faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. L'hypothèse était que les structures juridiques qui contenaient au moins une société non appariée satisfaisaient aux conditions d'inclusion dans la base de sondage. Sinon, les structures juridiques et, par conséquent, les entités juridiques qu'elles contenaient faisaient automatiquement partie de la base de sondage si au moins une de leurs sociétés satisfaisait à la règle du seuil limite.

2.3.3 Acquisition de données de sélection et de données de contact

L'étape précédente a abouti à une liste par approximation de toutes les entités commerciales pouvant éventuellement faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. Les données de sélection et de contact décrites dans la partie 2.1 qui n'étaient pas déjà dans la base de sondage étaient disponibles du RF. La base de sondage et le RF se recourent en partie parce que la majorité des unités de la partie de la base de sondage représentant les sociétés et une plus faible proportion des unités de la partie de la base de sondage représentant les particuliers étaient des employeurs. Il a fallu apparié les données de la base de sondage et les données du RF pour pouvoir ajouter à la base de sondage les données tirées du RF quand on constatait que des unités figuraient dans les deux sources. Autrement dit, il a fallu trouver quelles unités figuraient dans les deux sources.

toutefois justifiables puisqu'on pourrait par la suite corriger la base de sondage à l'aide des données obtenues du processus d'établissement des profils. La façon dont on a procédé pour créer la base de sondage est décrite en termes simples dans les paragraphes qui suivent.

La création de la base de sondage comportait trois étapes que nous allons décrire plus en détail un peu plus loin.

- i. Construire une liste de toutes les unités pouvant éventuellement faire partie de la base de sondage.
- ii. Déterminer quelles unités remplissaient les conditions requises pour faire partie de la base de sondage.
- iii. Obtenir des données de sélection et de contact.

2.3.1 Création des unités susceptibles de faire partie de la base de sondage

Les unités de sondage ont été constituées en groupant les entités juridiques pour créer des entités commerciales de la façon suivante. Les entités juridiques ont d'abord été groupées en structures juridiques. Une structure juridique se composait de l'ensemble des entités juridiques liées entre elles par la propriété de plus de 50 % des titres de propriété. Les relations comportant des entités juridiques étrangères étaient acceptées seulement si l'entité juridique étrangère possédait une entité juridique canadienne ou appartenait à une entité juridique canadienne. Dans les cas où une entité étrangère possédait plus d'une entité canadienne, la structure juridique a été divisée en autant d'entités commerciales qu'il y avait d'entités canadiennes appartenant directement à l'entité étrangère. Ainsi, un contact en vue de l'établissement d'un profil pouvait être pris avec le propriétaire canadien qui possède le contrôle de chaque entité commerciale résultante. Des exemples de cette façon de procéder sont donnés dans la Figure 1.

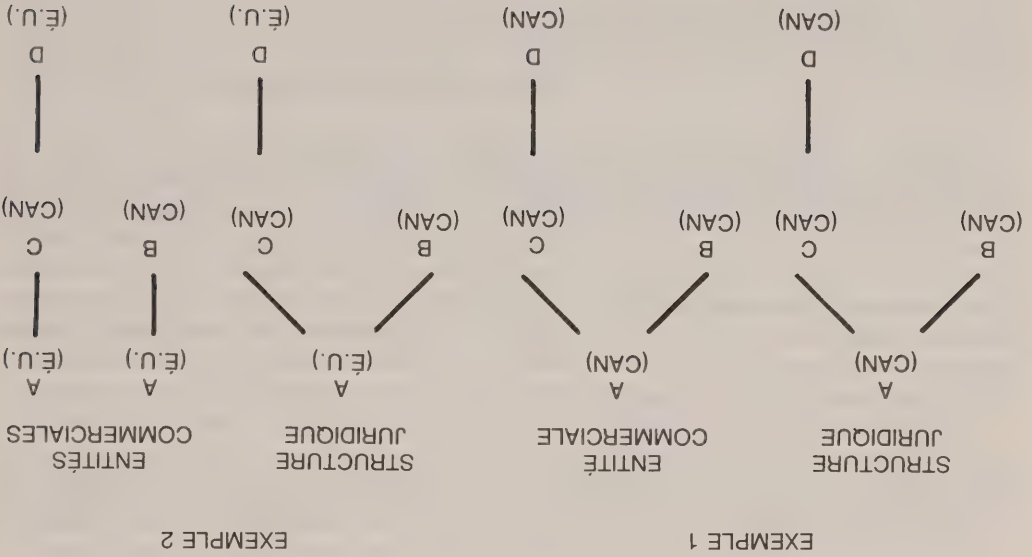


Figure 1. Définition des entités commerciales

bases de sondage. Les seules sources de données qui pouvaient être considérées étaient des listes de toutes les entités juridiques susceptibles d'entrer dans le champ des unités qui pouvaient être contactées en vue de l'établissement d'un profil et qui pouvaient produire au moins en partie certains des éléments d'information requis. Les sources de données retenues qui pouvaient être intégrées au moins partiellement par informatique étaient les suivantes:

- i. La Base de données sur les liens de parenté entre firmes (LPF), qui est une liste de toutes les entités juridiques ayant des activités au Canada et appartenant à une entité juridique étrangère ou canadienne, et ses propriétaires. Les entités juridiques étrangères sont couvertes jusqu'au niveau qu'il faut pour remonter jusqu'au propriétaire qui en dernière analyse possède le contrôle.
- ii. Le Registre des entreprises (RE) actuel, qui est principalement une liste de toutes les entités juridiques qui sont des employeurs. C'est dans ce registre qu'on peut trouver le nombre de locaux d'affaires d'une entité juridique, les données de contact (adresse et modalités de déclaration) utilisées pour les enquêtes et les industries dans lesquelles l'entité juridique est active.
- iii. La Base des unités de l'impôt sur le revenu des sociétés (IRS), qui est une liste de toutes les entités juridiques qui ont produit une déclaration d'impôt sur le revenu des sociétés auprès de Revenu Canada-impôt au cours d'une année donnée. L'activité principale, l'endroit où se trouve le siège social et le revenu de l'exercice financier sont indiqués dans cette source de données.
- iv. La Base des unités de l'impôt sur le revenu des particuliers (IRP), qui est une liste de tous les particuliers qui ont produit une déclaration d'impôt sur le revenu auprès de Revenu Canada-impôt au cours d'une année donnée. Les particuliers qui déclarent un revenu provenant d'un travail autonome sur leur déclaration sont des entités juridiques au sens des enquêtes économiques effectuées par Statistique Canada. On peut obtenir de cette base des données sur l'activité principale et des données de contact pour chaque particulier qui déclare un revenu provenant d'un travail autonome.

Les deux sources de données de l'impôt sur le revenu (IRS et IRP) sont des fichiers de données administratives. Les données administratives reçues tous les mois de Revenu Canada-impôt concernant les retenues sur la paye d'un employeur sont utilisées pour mettre à jour le RE. La source de données LPF est un fichier de réponses à une enquête par recensement. Aucune de ces sources de données ne couvre complètement l'univers des unités ni ne fournit tous les éléments d'information requis. La seule façon de couvrir tout l'univers est de les combiner. Cela vaut aussi pour certains éléments d'information requis tandis que pour les autres éléments, plus d'une source peuvent les fournir. La stratégie utilisée pour combiner ces sources de données afin d'obtenir le meilleur taux de couverture possible et la meilleure qualité des données possible est présentée dans la prochaine partie.

Une cinquième source, l'enquête sur le relevé trimestriel des états financiers, a fourni de l'information sur les entités juridiques qui préparent des états financiers consolidés. On a utilisé cette source pour manuellement affiner les entités commerciales sur la base de sondage.

2.3 Procédures de création de la base de sondage

Le problème que pose la création de la base de sondage servant aux contacts à prendre en vue de l'établissement des profils initiaux est d'intégrer quatre sources de données qui n'ont pas été conçues pour les mêmes objectifs et qui n'ont jamais été intégrées dans cette mesure auparavant. Cette difficulté est commune à tous les utilisateurs de données administratives. La tâche était encore compliquée par le fait que c'était la première fois qu'on appliquait en même temps autant de concepts établis pour la BDRC.

Des contraintes de temps et de ressources ont obligé l'équipe chargée du projet à faire certaines hypothèses au moment de créer la base de sondage. Les hypothèses formulées étaient

L'étape suivante a été de déterminer quelles unités de sondage composeraient la base de sondage et quelles données il fallait pour chacune. La base dont les entités commerciales seraient tirées pour un premier contact en vue de l'établissement d'un profil initial et à partir de laquelle on pourrait produire une première image de l'entité commerciale devait contenir toutes les entités commerciales pouvant être contactées.

Les entités peuvent être contactées en vue de l'établissement d'un profil si elles remplissent les conditions requises pour faire partie de la BDRCC. Les critères déterminant si une entité commerciale doit ou non être incluse dans la BDRCC sont appliqués à la structure juridique, qui décrit comment l'entité commerciale est organisée juridiquement.

Les entités juridiques peuvent faire partie de la partie intégrée de deux façons. Premièrement, si l'entité commerciale se compose d'une seule entité juridique, celle-ci entrera dans la partie intégrée si son revenu au cours de l'exercice financier considéré est supérieur à une valeur prédéterminée. Cette limite dépend de l'activité principale de l'entité juridique et de l'endroit où se trouve son siège social. Deuxièmement, si la structure juridique comprend plus d'une entité juridique, les entités juridiques feront toutes partie globalement de la partie intégrée si au moins une d'entre elles a un revenu supérieur à la limite établie pour ce genre d'entité.

Aussi, pour déterminer quelles entités commerciales pouvaient être contactées, il a fallu recueillir les renseignements suivants pour chaque entité juridique.

- i. Les relations de propriété entre les entités juridiques.
 - ii. Le revenu de l'exercice financier considéré, l'activité principale et l'endroit où se trouve le siège social.
- Pour ce qui est des entités commerciales qui remplissaient les conditions requises pour faire partie de la base de sondage et donc pour être contactées en vue de l'établissement d'un profil initial, il fallait des renseignements pour pouvoir les choisir et les contacter. Pour les choisir, il fallait les renseignements suivants:

- i. La liste de toutes les industries dans lesquelles l'entité commerciale avait des activités, pour pouvoir contacter d'abord les entités du secteur du commerce de gros et/ou du commerce de détail. Les enquêtes auprès de ces entités ont nécessité, avant les autres enquêtes, un ensemble d'entités statistiques produit à partir de contacts en vue de l'établissement de profils.
- ii. Le nombre de locaux d'affaires de toutes les entités commerciales comprenant une seule entité juridique ou deux entités juridiques si le propriétaire est un étranger. Cet élément d'information a déterminé le type de contact à prendre en vue de l'établissement de profils, qui pouvait être une interview téléphonique réalisée par le personnel d'un bureau régional ou une interview sur place réalisée par le personnel du bureau central ou d'un bureau régional.

- iii. La province où siège la société canadienne propriétaire principale. La province a été utilisée pour répartir entre les bureaux régionaux selon leur capacité la charge de travail en termes des contacts à prendre en vue de l'établissement de profils.
- Pour contacter les entités commerciales, il fallait le nom et l'adresse de l'entité juridique à la tête (excluant les propriétaires étrangers) de l'entité commerciale. Il était souhaitable de disposer en plus des données de contact et des données sur toute modalité de déclaration spéciale ayant été utilisées dans des enquêtes récentes.

2.2 Sources de données

Les sources de données qui pouvaient être utilisées étaient limitées principalement par le champ des unités que devait couvrir la base de sondage. Cette condition éliminait les listes d'éléments d'échantillons et beaucoup de listes propres à des industries précises comme des

du bureau central ou d'un bureau régional par interview sur place. Les données relatives aux autres entités seront recueillies par interview téléphonique. Les entités seront contactées une fois tous les deux ans ou plus souvent, selon la rapidité avec laquelle leurs structures changent.

La méthode d'établissement cyclique des profils, par laquelle les entités commerciales sont contactées périodiquement, est une des méthodes qui seront utilisées pour mettre à jour la FI de la BDR. On utilisera également des renseignements tirés d'enquêtes et des données de sources administratives.

La conception et la construction de la BDR s'échelonne sur une période de trois ans qui devrait aboutir à la création d'une base de données à intégrer dans les programmes des enquêtes. Lors de la mise en application de cette base, la plupart des données relatives à la partie intégrée de la BDR auront été obtenues au moyen du processus d'établissement des profils qui a commencé en avril 1986. Mais à ce moment-là, il n'y avait pas de liste unique d'entités commerciales dont on pouvait établir le profil.

Les données administratives ont joué un rôle important dans l'amorce du processus d'établissement des profils. Statistique Canada s'en est servi comme point de départ pour construire son registre des entités commerciales. Une liste des entités commerciales pouvant faire l'objet d'un profil initial a été dressée à partir de sources de données administratives. La partie 2 du présent document décrit comment cela s'est passé. Une description des sources de données utilisées pour construire la base de données de la BDR suit dans la partie 2.2, tandis que la partie 2.3 montre comment on a déterminé l'unité de sondage et comment on a combiné les diverses sources de données pour construire la base de sondage.

La partie 3 décrit comment les données administratives seront utilisées pour détecter les changements qui se seraient produits dans une entité commerciale et pour lancer ensuite le processus de mise à jour des profils. Sont ensuite présentés les résultats d'une étude de simulation effectuée pour quantifier le degré d'utilisation proposé des sources de données administratives. Enfin, le document se termine par une analyse de certains points soulevés dans cette étude.

2. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DE PROFILS INITIAUX

2.1 Éléments nécessaires à la création de la base de sondage

La première étape de la construction de la base de sondage devant servir à l'établissement des profils initiaux a été de définir l'unité de sondage. L'unité idéale aurait été l'entité commerciale. Toutefois, cette entité ne pouvait être obtenue à partir de sources ni intérieures ni extérieures à Statistique Canada. Les seules entités qu'il nous était possible d'obtenir étaient essentiellement des entités juridiques. Il a donc fallu regrouper les entités juridiques pour reproduire approximativement des entités commerciales. L'unité de sondage a été définie comme un agrégat d'entités juridiques assujetties aux contraintes suivantes.

- i. La définition de l'entité commerciale suppose que cette dernière englobe toutes les entités juridiques liées entre elles par des liens de propriété. La propriété est définie comme étant la possession de plus de 50% des actions avec droit de vote d'une entité légale. Le groupe-ment d'entités juridiques par cette forme de propriété est limité à un seul niveau de propriété étrangère à l'extérieur du Canada.
- ii. Il faut qu'une seule entité juridique canadienne ait la propriété de toutes les autres entités juridiques canadiennes faisant partie de l'entité commerciale. Cette condition est indispensable parce que les contacts en vue de l'établissement des profils avec l'entité commerciale pouvaient être pris au Canada seulement.

du fait que cette économie est dominée par un petit nombre de grandes entités commerciales qui représentent le gros de l'activité dans cette économie. La BDRCC est divisée en deux parties pour traduire cette dichotomie.

Une des composantes, la partie intégrée (PI), couvre le petit nombre d'entités commerciales importantes par leur taille ou selon d'autres critères, tandis que l'autre composante, la partie non intégrée (PNI), couvre les autres entités commerciales, c'est-à-dire le grand nombre de petites entités. Les entités de la première composante sont plus complexes. Aussi faut-il un certain effort pour déterminer, dans une entité commerciale complexe, quels éléments peuvent être intéressants pour une enquête donnée.

La partie intégrée (PI) de la BDRCC est un moyen de représentation de la structure complexe des entités commerciales à l'aide d'un modèle d'information. Le modèle se compose de cinq structures liées entre elles qui décrivent une entité commerciale. Ces structures permettent de définir avec précision les populations d'enquête. Les cinq structures sont les suivantes.

- i. *La structure juridique*, qui décrit comment l'entité commerciale est organisée juridiquement. Elle représente les entités juridiques et les relations de propriété et de contrôle qu'elles ont entre elles. Entre autres exemples d'entités juridiques, notons les entreprises constituées en société en vertu de chartes fédérales ou provinciales.

- ii. *La structure opérationnelle*, qui décrit comment l'entité commerciale fonctionne et comment elle organise son système comptable. Elle se compose des entités exploitantes. C'est elle qui organise et contrôle la production des biens et/ou des services. Il s'agit d'un moyen de structurer l'entité commerciale de la façon dont cette dernière se perçoit. Entre autres exemples d'entités exploitantes, notons les divisions, les centres de profit et les usines.
- iii. *La structure statistique*, qui représente les entités statistiques classées selon un ordre hiérarchique. Les entités statistiques sont constituées à partir de la structure opérationnelle correspondante suivant les unités de la structure opérationnelle pour lesquelles un ensemble particulier de données est tenu à jour.

- iv. *La structure déclarante*, qui représente les modalités de déclaration définies pour chacune des entités statistiques choisies, par enquête. Les données du système comptable de l'entité commerciale sont communiquées par les entités déclarantes.
- v. *La structure administrative*, qui contient des données administratives comme les données fiscales recueillies auprès des entités juridiques et les données des comptes de retenue sur la paye recueillies auprès des entités exploitantes.

Les entités des structures statistiques et déclarantes sont produites par le Bureau à des fins de collecte, de vérification, d'estimation et de totalisation de données économiques. Les entités des trois autres structures sont définies par les organismes externes à Statistiques Canada. Le processus complexe de délimitation des frontières de chacune des entités commerciales et de détermination de ses cinq structures et des relations existant entre ces structures est appelé «établissement de profils». Cette représentation de l'entité commerciale comme réseau est le «profil». Les données servant à construire un profil sont obtenues par contact avec l'entité commerciale ou avec une de ses composantes. Les éléments d'information sur la structure juridique et sur la structure opérationnelle des entités commerciales ainsi que certains éléments d'information sur leur structure administrative sont obtenues ou révisées et mis à jour au cours de l'interview. La structure statistique est ensuite produite ou mise à jour automatiquement à partir de la nouvelle structure opérationnelle. Finalement, des entités déclarantes implicites sont créées pour chaque entité statistique nouvellement choisie à l'aide des données de certaines zones tirées des structures juridique, opérationnelle ou administrative. Ces entités peuvent par la suite être mises à jour au moyen des renseignements obtenus à l'occasion du premier contact avec les répondants ou au moyen d'arrangements spéciaux négociés avec les répondants. Le type de contact qui est pris en vue de l'établissement des profils dépend de la complexité des entités et de toute modalité de déclaration spéciale. Pour ce qui est des données relatives aux entités les plus complexes et les plus importantes, elles seront recueillies par le personnel

Utilisation de données administratives pour l'établissement des profils initiaux et ultérieurs des entités économiques

COLLEEN CLARK et ROBERT LUSSIER¹

RÉSUMÉ

Statistique Canada s'emploie actuellement à remanier son registre central des entités économiques. Dans le nouveau registre, chaque entité économique est considérée comme un réseau d'entités juridiques et exploitantes dont les caractéristiques permettent de déterminer des entités statistiques. On obtient l'image de ce réseau, c'est-à-dire le profil par le processus dit d'«établissement des profils», qui suppose des contacts avec l'entité économique. En 1986, on s'est servi d'une liste de toutes les entités avec lesquelles il fallait entrer en contact afin d'obtenir les profils permettant de constituer le nouveau registre. Pour dresser cette liste, on a eu recours à des données administratives. À l'avenir, les données administratives serviront de source de renseignements sur les changements qui auront pu se produire dans les entités économiques. Elles pourront donc être utilisées comme source de mise-à-jour directe ou comme signal qu'on doit réviser la structure d'une entité. L'article porte d'abord sur les objectifs du processus d'établissement des profils. On présente ensuite les procédures de construction de la base de sondage servant au processus d'établissement des profils initiaux à l'aide de plusieurs sources de données administratives. Ces procédures comprennent l'application de concepts, la détection des cas de chevauchement entre les sources et l'évaluation de la qualité des données. On examine ensuite le rôle des données administratives comme source de renseignements sur les changements qui peuvent s'être produits dans les entités économiques et comme source de données sur lesquelles on peut s'appuyer pour demander de vérifier les profils. Suit une analyse des résultats d'une étude de simulation visant à évaluer ce rôle. L'exposé s'achève par une série de questions sur la méthodologie relative à l'utilisation de données administratives en vue de l'établissement de profils de mise à jour.

MOTS CLÉS: Données administratives; registre central; profil.

1. INTRODUCTION

Statistique Canada s'emploie actuellement à réorganiser son programme d'enquêtes économiques. Dans le nouveau programme, on utilisera davantage les données administratives. Ces dernières feront partie intégrante d'une Base de Données du Registre Central (BDRC) d'où les enquêtes économiques tireront leur échantillons. Les données administratives serviront aussi à mettre à jour la BDRC. Cet aspect et d'autres aspects de la stratégie de réorganisation sont présentés dans Colledge et Lussier (1985). Certains résultats de la mise en pratique de la stratégie sont contenus dans Colledge (1987). Une des premières étapes a été de définir les unités de la BDRC. L'unité fondamentale est l'entité commerciale. Statistique Canada (voir Statistique Canada, 1987) définit une entité commerciale comme: «un agent économique ayant la responsabilité et le pouvoir d'affecter des ressources à la production de biens et/ou de service et, de ce fait, de gérer et de contrôler la façon de recevoir et d'employer les recettes, d'accumuler des avoirs, d'emprunter et de prêter des capitaux et de tenir à jour des états financiers complets de ses activités» (traduction). La Base de Données du Registre Central actuellement élaborée par Statistique Canada est une entreprise de représentation de la structure de l'économie canadienne. Elle tient compte

¹ Colleen Clark, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-C1, Immeuble Jean Talon, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6; Robert Lussier, Division des méthodes d'enquêtes-entreprises, 11-M, Immeuble R.H. Coats, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

- MARQUIS, K. (1986). Discussion de «Correlates of Reinterview Inconsistency in the Current Population Survey». *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, 235-240.
- MARQUIS, K. (1978). Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations. Technical Report R-2319-HEW, The Rand Corporation.
- MARQUIS, K., MARQUIS, S., et POLICH, M. (1986). Response bias and reliability in sensitive topic surveys. *Journal of the American Statistical Association*, 81, 381-389.
- MOORE, J., et KASPRZYK, D. (1984). Month-to-month reciprocity turnover in the ISDP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 726-731.
- NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). An Overview of the Survey of Income and Program Participation, Update I. SIPP Working Paper Series, n° 8401, U.S. Bureau of the Census.
- NETER, J., MAYNES, S., et RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- NETER, J., et WAKSBERG, J. (1966). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- SATER, D. (1986). SSN Response Rates and Results of SSN Validation Improvement Operation. Note de service du U.S. Bureau of the Census.
- STASNY, E., et FIENBERG, S. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in the Labor Force Statistics*, U.S. Departments of Commerce and Labor, 25-39.

REMERCIEMENTS

Bon nombre de personnes ont contribué à la réalisation de l'étude de vérification des dossiers SIPP. Bien que nous ne puissions pas nommer ici toutes celles qui ont pris part au projet, nous désirons témoigner notre reconnaissance à Jeanette Robinson, qui a préparé la multitude de fichiers des dossiers administratifs en vue de l'appariement, à Elaine Fansler, qui a préparé et exécuté un nombre incalculable de passages en machine pour l'appariement, à Bill LaPlant, que nous avons consulté en raison de ses vastes connaissances techniques concernant le programme d'appariement du Census Bureau et les programmes qui s'y rattachent, à Chris Dyke, qui s'est montré infatigable dans ses efforts pour assurer le bon déroulement du programme d'appariement dans le nouveau système informatique, et à Dan Kasprzyk, pour l'appui constant et patient qu'il a témoigné à l'égard de notre projet. Le présent document a aussi fait l'objet d'un examen sérieux de la part du rédacteur et de deux arbitres anonymes de *Techniques d'enquête*; nous désirons exprimer notre reconnaissance pour les suggestions et les commentaires qu'ils nous ont offerts.

BIBLIOGRAPHIE

BAILLAR, B. (1968), Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.

BURKHEAD, D., et CODER, J. (1985). Gross changes in income reciprocity from the Survey of Income and Program Participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-356.

COX, B., et IACHAN, R. (1987). A comparison of household and provider reports of medical conditions. *Journal of the American Statistical Association*, 82, 1013-1018.

DAVID, M. (1983). *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program*. New York: Social Science Research Council.

FEATHER, J. (1972). *A Response Record Discrepancy Study*. Saskatoon, Saskatchewan: Université de la Saskatchewan.

FELLEGI, I., et SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

FULLER, W., et TIN, C.C. (1986). Response error models for changes in multinomial variables. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, 425-441.

HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 210-215.

JARO, M. (1985). Current record linkage research. Communication présentée au Census Advisory Committee de l'American Statistical Association, U.S. Bureau of the Census, 25 avril.

JÖRESKOG, K., et SÖRBOM, D. (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*. Mooresville, Indiana: Scientific Software, Inc.

KOONS, D. (1973). Quality control and measurement of nonsampling error in the Health Interview Survey. *Vital and Health Statistics*, série 2, 54.

LAPLANT, W. (1987). Maintenance Manual for the Generalized Record Linkage Program Generator (GENLINK) SRD Program Generator System, Statistical Research Division, document interne, U.S. Bureau of the Census.

LUENBERGER, D. (1984). *Linear and Nonlinear Programming*. Reading, Massachusetts: Addison Wesley.

Fournis ici indiquent que le fait d'exclure les personnes non présentes durant les huit mois des périodes de référence des vagues 1 et 2 n'influe pas sur les tendances montrées aux figures 3 et 4. En raison du faible nombre de cas et du fait que l'échantillon du Wisconsin n'est pas représentatif, nous n'offrons pas de statistiques obtenues par déduction pour cette série d'illustrations.

Dans les figures, les rectangles hachurés indiquent le nombre de changements observés selon les dossiers administratifs et les rectangles non hachurés, le nombre de changements relevés selon la SIPP. S'il y a trop de transitions au point de jonction d'après la SIPP, le rectangle non hachuré domine le rectangle hachuré pour les comparaisons relatives à la jonction. Si un trop petit nombre de transitions sont relevées dans le cadre de la SIPP pour les mois visés par une interview, le rectangle hachuré est censé être moins important que le rectangle hachuré pour les comparaisons portant la « vague 1 » et la « vague 2 ». Et, si les interviews de la SIPP donne lieu approximativement au bon nombre de rapports de transition, les rectangles hachuré et non hachuré doivent être à peu près de la même taille pour les comparaisons ayant trait à la moyenne pour tous les mois.

La figure 3 présente la fréquence moyenne, d'un mois à un autre, des transitions relatives à la participation à l'AFDC au Wisconsin dans les vagues 1 et 2 pour les deux sources de données, et compare ces données à celles qui sont observées au point de jonction des vagues 1 et 2. Le problème du « biais attribuable au point de jonction » observé au moyen des données de la SIPP est passablement apparent, la fréquence des transitions au point de jonction étant supérieure à la moyenne pour l'une et l'autre des interviews. Bien que les différences absolues pour cette taille d'échantillon soient faibles, les données des dossiers laissent supposer que le biais imputable au point de jonction pour l'AFDC résulte d'une combinaison d'un trop grand nombre de transitions déclarées au point de jonction et d'un trop petit nombre de transitions relevées entre les autres mois visés par l'interview. De plus, les dernières colonnes de la figure 3 laissent supposer un réel problème de déclaration par défaut des transitions pour l'AFDC selon la SIPP ainsi qu'un problème de répartition en fonction du temps.

Les résultats concernant le programme des bons alimentaires au Wisconsin sont résumés à la figure 4, où l'effet du biais attribuable au point de jonction est encore plus évident. Là encore, les données des dossiers administratifs laissent supposer que les transitions relatives à une interview ont tendance à être sous-estimées par comparaison aux données de la SIPP. En outre, dans le cas présent, le contraste entre les données de l'enquête et celles des dossiers administratifs montre encore plus clairement que les transitions observées au point de jonction selon la SIPP sont très surestimées. Toutefois, contrairement aux résultats de la participation à l'AFDC, les données de l'enquête et des dossiers indiquent à peu près le même nombre de transitions dans l'ensemble, ce qui laisse supposer qu'il s'agit d'un simple problème de répartition en fonction du temps et non d'un problème réel de déclaration par défaut.

7. CONCLUSIONS

Le long processus d'appariement et de préparation des dossiers est maintenant terminé et nous amorçons notre analyse de ce riche ensemble de données. Toutefois, nous avons déjà montré, à l'aide des seules données initiales présentées ici, de quelle manière les résultats de la vérification des dossiers peuvent nous aider à mieux comprendre les questions importantes se rapportant aux erreurs de mesure, dans le cas présent le biais attribuable au point de jonction de la SIPP. Il y a beaucoup d'autres tests à faire et beaucoup d'hypothèses à explorer avant qu'on puisse tirer des conclusions définitives au sujet de la nature des erreurs de mesure de la SIPP et de leurs causes probables. Nous sommes persuadés que l'étude de vérification des enregistrements SIPP nous permettra de mieux comprendre l'importance et la nature de ces erreurs d'enquête et nous aidera peut-être à déterminer leur cause.

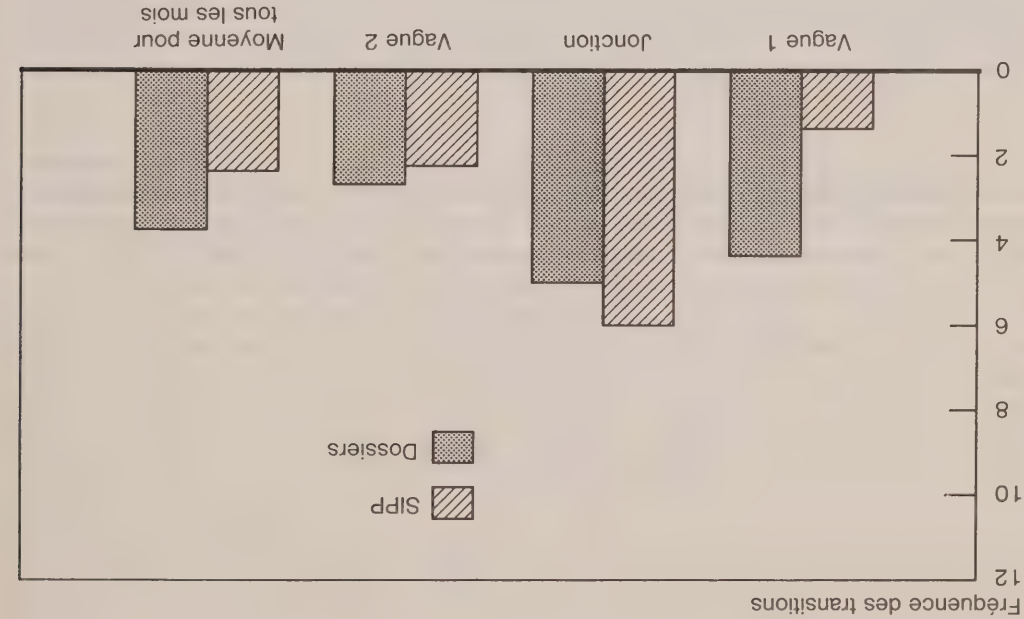


Figure 3 : Transitions d'un mois à un autre relatives à la participation à l'AFDC : comparaison de la fréquence des transitions à la jonction avec la fréquence moyenne dans les vagues 1 et 2, et moyenne globale pour tous les mois selon la SIPP et les dossiers administratifs.

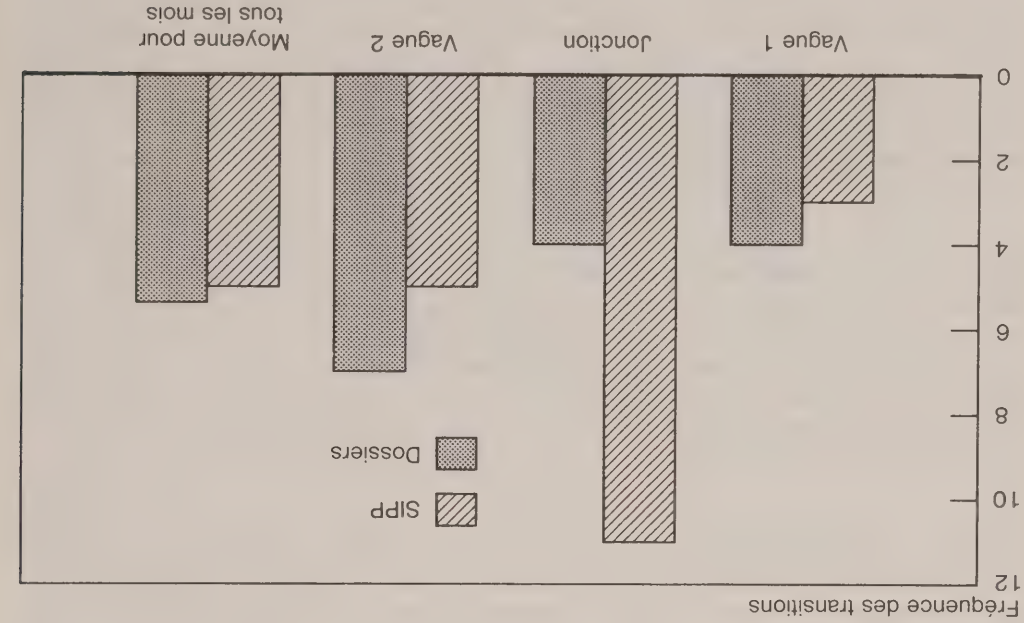


Figure 4 : Transitions d'un mois à un autre relatives à la participation au programme de bons alimentaires : comparaison de la fréquence des transitions à la jonction avec la fréquence moyenne dans les vagues 1 et 2, et moyenne globale pour tous les mois selon la SIPP et les dossiers administratifs.

types d'erreurs de réponse, de non-réponse ou d'application des directives qui pourraient être à l'origine de ce modèle et quel ensemble d'estimations de la transition est le plus approprié. En analysant le problème en fonction des données administratives, nous espérons pouvoir déterminer de façon plus sûre dans quelle mesure les erreurs de réponse et de non-réponse contribuent à créer le modèle observé.

Il est possible qu'il y ait un lien entre le biais attribuable au point de jonction et le phénomène mieux connu selon lequel la variance des erreurs de mesure a tendance à gonfler les estimations de changement brut ou à sous-estimer la stabilité. Selon des ouvrages récents (Fuller et Tin, 1986), il existe plusieurs solutions à ce problème. Nous prévoyons entreprendre une étude empirique des effets des erreurs de mesure sur les estimations de la transition afin de déterminer s'il est possible, par exemple, de corriger les erreurs de réponse en nous fondant sur les estimations établies à la suite des réentrevues.

Enfin, nous avons mentionné précédemment qu'il pourrait être difficile d'obtenir des estimations non faussées des erreurs si les dossiers contiennent aussi des erreurs. Nous voulons estimer, à l'aide des mesures de réinterview (qui identifient l'estimation de Var *e*), la variance des erreurs dans les dossiers (Var *u*). Toutefois, nous continuons de supposer que les dossiers sont sans biais.

6. PREMIÈRES CONCLUSIONS

Pour illustrer notre approche, nous examinons la question du «point de jonction» à l'aide des données relatives à deux programmes de transfert gouvernemental d'un État. Rappelons que le problème du point de jonction résulte du fait que les rapports mensuels d'interview indiquent un plus grand nombre de changements relatifs à la participation aux programmes entre les mois visés par des interviews distinctes qu'entre les autres mois (visés par la même interview). Grâce aux données de dossiers administratifs, nous sommes en mesure de commencer à répondre aux principales questions concernant la qualité des estimations des changements dans le cadre de la SIPP: Y a-t-il trop de changements déclarés au point de jonction? Y a-t-il trop peu de changements déclarés pour les autres mois? Est-ce que la SIPP déclare le nombre exact de changements pour tous les mois d'enquête, mais répartit ces changements de façon incorrecte?

Les figures 3 et 4 contiennent les résultats de nos analyses initiales du biais attribuable au point de jonction. Les données de ces analyses initiales proviennent des fichiers appariés/fusionnés de la SIPP et des dossiers administratifs pour le programme d'aide financière aux familles à faible revenu avec enfants à charge et le programme de bons alimentaires de l'État du Wisconsin. Dans le cadre de la SIPP, il y avait un échantillon de 1,632 personnes admissibles au Wisconsin dans la vague 1 du panel de 1984 de la SIPP. De ce total, 92 (6%) ont refusé de déclarer leur SSN et ont conséquemment été exclues de l'appariement des dossiers administratifs et des analyses des erreurs de réponse. De plus, l'échantillon du Wisconsin fait partie d'un échantillon national et n'est pas nécessairement représentatif de cet État.

Dans le cadre de la SIPP, on suppose que toutes les personnes échantillonnées faisant partie de la vague 1 sont des personnes admissibles qui appartenaient au même ménage durant tous les mois de la période de référence de la vague 1, et que personne autre que les personnes admissibles à l'interview de la vague 1 était membre du ménage au cours des quatre mois précédents. Ainsi, les estimations de la transition d'un mois à l'autre pour la vague 1 ont été établies à partir d'une base constante de 1,540 répondants (1,632 - 92). Toutefois, dans la vague 2, la fluidité de la composition des ménages est reconnue et donne lieu à des bases de répondants qui varient légèrement d'une période de deux mois à une autre, y compris la période de jonction. Dans les données ci-après, le nombre de personnes admissibles pour les deux mois de la période de jonction a atteint 1,517; dans la vague 2, les bases de répondants pour les trois périodes de deux mois s'élèvent à 1,522, 1,531 et 1,532. Les résultats d'analyses distinctes non

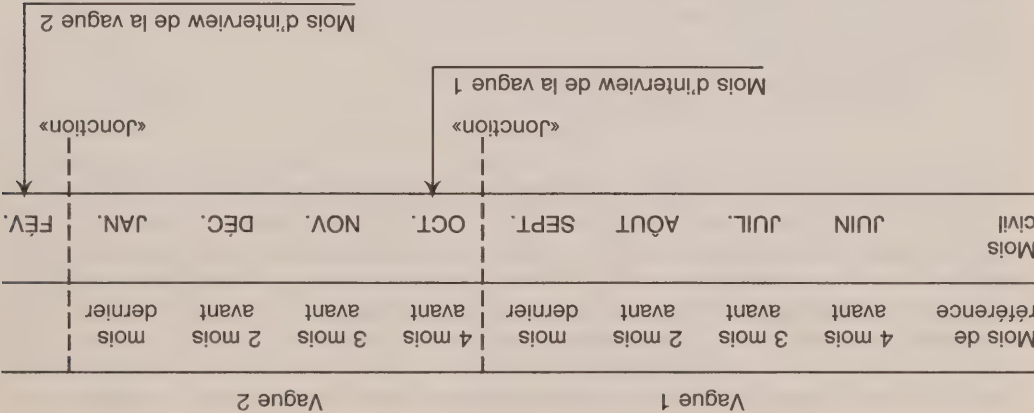


Figure 2: Périodes d'enquête prévus pour le groupe de renouvellement 1 dans le cadre de la SIPP et indiquant les mois de référence, les mois civils, les mois d'interview et le point de «jonction».

Pour connaître l'applicabilité de l'hypothèse de télescopage interne (selon laquelle les personnes n'oublient pas certains faits, mais croient que ces faits sont survenus à une date moins éloignée que la date réelle), nous déterminerons si l'erreur systématique de réponse relative aux premiers mois de la période de référence (juin et juillet pour la vague 1 et octobre et novembre pour la vague 2) est négative et si elle est positive pour les derniers mois (août et septembre, décembre et janvier), et si la somme de ces deux erreurs correspond à zéro.

Nous avons l'intention de vérifier l'hypothèse qui s'applique aux interviews portant sur une période limitée et selon laquelle les répondants déclarent les faits qui se sont passés il y a longtemps comme ayant eu lieu au cours d'une période de référence non limitée par une interview précédente (juin à septembre), mais qu'une telle chose ne se produit pas lorsque les périodes de référence sont délimitées par une interview antérieure (dans le cas présent, octobre à janvier). Pour examiner l'hypothèse au sujet des défaillances de mémoire (voulant que la probabilité d'oublier un fait augmente avec le temps), nous déterminerons si l'erreur systématique de réponse à une valeur négative plus élevée au cours des premiers mois de chaque période de référence qu'au cours des derniers mois.

Les hypothèses relatives à la durée du temps passé dans l'échantillon et aux groupes de renouvellement supposent que les erreurs de réponse seront plus importantes à la deuxième interview qu'à la première, après avoir corrigé les effets résultant des variations saisonnières. Nous étudierons cette hypothèse et si nous constatons qu'elle est juste, nous examinerons certaines des théories avancées par des spécialistes qui expliquent pourquoi un tel phénomène peut se produire. Est-ce que les répondants de la première interview qui participent à la seconde interview sont différents, comme le laissent entendre Stasny et Fienberg (1985), ou est-ce que la qualité des déclarations des répondants diminue à la deuxième interview, comme le prévoit l'hypothèse du conditionnement de Neter et Wakseberg (1966)?

Nous n'avons pas encore déterminé dans quelle mesure ces problèmes classiques des enquêtes longitudinales se posent pour la SIPP. Toutefois, un des problèmes connus à l'estimation des changements observés d'un mois à un autre dans la participation aux programmes (Burkhead et Coder, 1985). Plus précisément, un plus grand nombre de changements dans la participation aux programmes survient au «point de jonction» entre les interviews (entre septembre et octobre, figure 2) qu'entre les mois visés par une interview (par ex., entre juin et juillet, ou entre août et août, ou entre août et septembre). Le Census Bureau n'a pas encore publié les estimations mensuelles de la transition relative à la participation aux programmes parce que ces estimations révèlent un modèle sur lequel l'erreur de mesure semble avoir influé considérablement. Moore et Kasprzyk (1984) ainsi que Hill (1987) se sont interrogés sur les

3. calcule un poids d'appariement composite relatif à toutes les paires d'enregistrements possibles dans le bloc;
4. dans le bloc, fait correspondre à chaque enregistrement d'un fichier un enregistrement apparié d'un autre fichier selon une formule qui maximise le poids composite total pour toutes les paires du bloc;
5. applique la technique décisionnelle de Fellegi-Sunter pour déterminer si une paire d'enregistrements est assortie, non assortie ou nécessite un autre examen; et
6. établit un fichier de correspondance se rapportant aux enregistrements appariés dans chaque fichier.

5. ANALYSE

L'objet de l'étude de la vérification des enregistrements est d'estimer certains paramètres des erreurs de mesure applicables aux échantillons de personnes, au contenu des données et aux mois d'enquête, et d'évaluer le rapport qui existe entre les erreurs mêmes et entre les erreurs et les variables qui reflètent les caractéristiques du plan de sondage. L'objectif général de l'étude est d'utiliser les données appariées en vue d'estimer pour chaque variable dichotomique relative à la participation:

1. l'erreur systématique de réponse (à l'aide du résultat de la différence de valeur des données de l'enquête et des enregistrements);
2. les variables explicatives de l'erreur systématique de réponse (à l'aide des techniques de régression logistique ou de régression des probits, ou peut-être des techniques LISREL fondées sur des matrices présentant des coefficients d'association à séries multiples et tétrachorique (Jöreskog et Sörbom 1984);
3. la variance des erreurs de réponse (par ex., établie à partir des résidus de la régression);
4. les conditions ou les groupes qui présentent des variances des erreurs de réponse très importantes et très petites; et
5. le genre et le degré de confusion relative aux programmes de transfert qui engendrent des erreurs de réponse (à l'aide de méthodes d'analyse de la structure de covariance comme LISREL).

(Nous estimerons les mêmes paramètres pour les déclarations des sommes d'argent reçues dans le cadre de chaque programme de transfert, mais nous n'avons pas encore choisi la méthode d'estimation de base.)

Les questions concernant les erreurs de mesure sont réparties en deux catégories: les questions qui s'appliquent à tous les mois d'enquête et celles pour lesquelles on compare les erreurs d'un mois d'enquête à un autre. La première catégorie comprend les estimations du nombre d'erreurs de réponse imputées aux répondants visés et aux enquêtes-substituts et celles qui sont attribuées aux intervieweurs. La deuxième catégorie comprend les erreurs qui résultent des enquêtes par panel et que l'on connaît bien, par exemple, les erreurs de télescopage, le biais dû à la durée de la présence dans l'échantillon, les défaillances de mémoire, le biais attribuable au groupe de renouvellement, *etc.*, c'est-à-dire celles qui sous-entendent que les erreurs de mesure différeront d'un mois d'enquête à un autre, toute autre chose demeurant constante par ailleurs. Nous ajoutons à cette liste ce que Hill (1987) a qualifié de biais du «point de jonction» des enquêtes longitudinales et dont nous parlerons ci-dessous.

Pour mieux comprendre les questions que nous désirons aborder au sujet des différentes périodes de déclaration, il faut se reporter à la figure 2 qui présente le calendrier des mois d'interview et des périodes de référence pour un groupe de renouvellement de répondants SIPP. La figure fait état de deux interviews. La première a lieu au début d'octobre et recueille des renseignements sur des faits survenus en septembre (le mois précédent), en août (deux mois avant), en juillet (trois mois avant) et en juin (quatre mois avant). De même, au cours de la deuxième interview qui est effectuée quatre mois plus tard, les données recueillies portent sur janvier, décembre, novembre et octobre. Nous nommons «jonction» la période de transition entre septembre et octobre parce que cette période se situe entre les périodes de référence visées par les deux interviews.

L'étude emploie des stratégies de blocage multiples et indépendantes pour chaque paire de fichiers à appairer, ce qui réduit la possibilité qu'une paire réellement assortie ne sera pas repérée par suite du groupage. Une des principales stratégies de groupage consiste à utiliser les trois premiers chiffres du code de zone de cinq chiffres du service postal américain et un code SOUNDEX de quatre caractères tiré du nom de famille de la personne qui fait partie de l'échantillon ou du bénéficiaire. Le code de zone est un indicateur géographique d'une division régionale d'un Etat qui, selon les experts en appariement du Census Bureau, ne comporte habituellement pas d'erreur. L'algorithme SOUNDEX est fréquemment utilisé pour créer un code de même longueur et de même structure à partir de chaînes de caractères en entrée de longueurs variables; lorsqu'il est utilisé à des fins de groupage, il présente un avantage du fait qu'il permet de réduire le nombre d'erreurs de groupage attribuables aux erreurs d'orthographe, mais pas de les supprimer complètement. Une deuxième stratégie de groupage est fondée sur les quatre derniers chiffres du SSN.

4.2.3 Poids d'appariement des zones de données

Compte tenu de certaines variations, les zones de données utilisées pour effectuer l'appariement des fichiers SIPP et des données administratives comprennent le numéro de voirie, le nom de la rue, le numéro d'appartement, la ville, le code de zone, le SSN, le sexe, la date de naissance, le nom de famille et le prénom. Nous savons intuitivement que certaines de ces zones sont plus valables que d'autres lorsqu'il s'agit de déterminer si une paire donnée d'enregistrements est assortie ou non; ainsi, la concordance de la variable «sexe» ne constitue pas, de toute évidence, un indice d'appariement réel aussi sûr que la concordance de la variable «SSN», par exemple. Dans leur présentation d'une théorie générale de couplage des enregistrements, Fellegi et Sunter (1969) discutent du calcul des poids qui tiennent compte des différents pouvoirs de discrimination des diverses zones de données et de quelle façon les règles de décision optimale utilisent les poids. Le personnel de recherche en matière de couplage des enregistrements du Census Bureau a élaboré des programmes en se fondant sur la méthode d'établissement des systèmes non linéaires de Newton (voir Luenberger, 1984) pour résoudre les équations de Fellegi-Sunter, et ces programmes sont utilisés dans le cadre de l'étude de vérification des enregistrements SIPP pour calculer les poids finals d'appariement.

4.2.4 Programme d'appariement automatisé

Le programme d'appariement automatisé du Census Bureau applique les calculs de Fellegi-Sunter à un ensemble défini par l'utilisateur de zones de données tirées de fichiers triés (groupés) selon les exigences de l'utilisateur. Pour chaque zone de données devant servir à l'appariement, l'utilisateur introduit les valeurs de départ des poids d'appariement, détermine les critères de concordance pour la comparaison (c.-à-d. si les zones doivent être parfaitement identiques pour que le programme d'appariement détermine qu'il y a concordance ou si une comparabilité approximative suffit), relève les inscriptions manquantes et précise comment il faut procéder dans ces cas (les inclure ou ne pas en tenir compte dans le calcul d'un poids d'appariement composite). L'utilisateur détermine les valeurs limites des poids composites pour les paires appariées et non appariées, et produit les codes appropriés du programme en COBOL en vue d'effectuer un appariement à l'aide de GENLINK, le générateur de programmes de couplage des enregistrements du Census Bureau (LaPlant 1987).

Autrement dit, le programme d'appariement:

1. parcourt chaque fichier de données afin de trouver des blocs d'enregistrements comparables, c'est-à-dire des enregistrements dont les éléments choisis du groupage concordent parfaitement;
2. compte le nombre d'enregistrements dans les blocs relevés pour s'assurer que la longueur des blocs de l'un ou l'autre des fichiers ne dépasse pas la longueur maximale prédéterminée;

4.2 Méthodes d'appariement automatisé du Censur Bureau

L'étude de vérification des enrégistrement est effectuée à l'aide de méthodes d'appariement automatisé qui sont fondées sur l'ouvrage théorique traitant du couplage des enrégistrement de Fellegi et de Sunter (1969). Le processus comporte des mesures discontinues multiples, mais fondamentalement, on en compte quatre:

1. uniformiser les zones communes de données des deux fichiers que le programme d'appariement examinera afin de déterminer si on peut appairer ou non une paire d'enregistrements;
2. trier les deux fichiers de façon à établir de petits sous-ensembles d'enregistrements (ou «blocs» d'enregistrements) qui sont constitués d'un nombre adéquat de paires que le programme d'appariement doit vérifier;
3. déterminer et quantifier l'utilité de chaque zone de données qui pourrait faire l'objet de l'appariement visant à relever les concordances parfaites; et
4. appliquer les algorithmes informatiques qui permettent d'effectuer l'appariement proprement dit des enrregistrements.

4.2.1 Uniformisation

Dans le cadre de l'étude de vérification des enrregistrements, tous les fichiers de données, c'est-à-dire les fichiers SIPP et ceux des dossiers administratifs, sont traités à l'aide d'un programme d'uniformisation des données sur l'adresse qui uniformise la structure des diverses composantes de l'adresse (par ex., le nom de la rue, le type de rue et son orientation, le nom de la ville, l'abréviation utilisée pour l'Etat, etc.) et analyse chaque composante d'une zone fixe de données. Plusieurs programmes ont été établis à cette fin. Nous utilisons le programme ZIPSTAN qui a été élaboré au Censur Bureau.

En plus des méthodes d'uniformisation qui s'appliquent à tous les fichiers de données, les zones individuelles de données de bon nombre de fichiers doivent être modifiées de façon que les fichiers présentent une même structure pour les besoins de l'appariement. Parmi les variables qui présentent un problème sur ce plan, nous retrouvons le sexe (qui peut être indiqué par un code alphabétique, «m» ou «f», ou numérique, «1» ou «2»); la date de naissance (qui peut être indiquée de différentes façons: «mm-jj-aa», «ss-aa-mm-jj» ou selon la date julienne); et le nom (il peut y avoir une seule zone ou des zones distinctes pour chaque composante). Nous établissons des programmes sur demande pour faire ce genre d'uniformisation.

4.2.2 Groupage d'enregistrements

Le groupage d'enregistrements, qui consiste à créer des sous-ensembles d'enregistrements que le programme d'appariement examinera en vue d'assortir des paires d'enregistrements (par ex., Jaro 1985), est une technique nécessaire lorsqu'il faut appairer des fichiers qui contiennent un grand nombre d'enregistrements. De toute évidence, il y aurait une probabilité maximale de faire des appariements parfaits si pour chaque enrégistrement d'un fichier, on parcourait l'autre fichier en entier pour trouver l'enrégistrement correspondant. Toutefois, il est impossible de procéder à de telles recherches dans des fichiers de cette taille. Grâce aux sous-ensembles d'enregistrements qui sont établis pour chaque fichier, il est possible de faire l'appariement; toutefois, certains enrregistrements sont alors exclus des sous-ensembles, ce qui accroît la probabilité que certains rapprochements exacts ne seront pas faits. Par conséquent, il faut s'assurer que les éléments groupés présentent suffisamment de variation pour permettre la répartition des fichiers en de nombreux (et plus petits) blocs, et que ces éléments permettent de discerner de façon rapide s'il y a ou non appariement, c'est-à-dire qu'ils concordent presque toujours lorsqu'une paire d'enregistrements est bien assortie et ne concordent presque jamais lorsque l'inverse se produit.

Comme il a été mentionné précédemment, les erreurs contenues dans les dossiers peuvent nuire aux études de vérification des enregistrements. Bien que plusieurs des fichiers de données administratives obtenus pour ce projet comportent quelques légères failles, seulement deux d'entre eux semblent pouvoir poser de sérieux problèmes sur le plan analytique: le fichier de l'indemnisation des accidents du travail de l'Etat de New York et le fichier des indemnités et pensions des anciens combattants. Chacun de ces fichiers présente un relevé incomplet des bénéficiaires. Le premier fichier ne tient pas compte d'un nombre inconnu de cas considérés comme «régles» (c.-à-d. de cas sur lesquels on s'était déjà prononcé et pour lesquels des paiements avaient déjà commencé à être versés par une entreprise d'assurance privée) au moment où la base de données a été établie il y a plusieurs années. Le deuxième fichier ne contient pas de données sur les bénéficiaires dont les prestations ont été envoyées à un établissement financier ou autre; ces bénéficiaires représentent environ un pour cent de l'ensemble des bénéficiaires. Les autres fichiers ne semblent pas présenter de problèmes de cet ordre.

L'écart entre la date d'émission du chèque et la date de réception constitue un problème inévitable qui touche tous les fichiers administratifs dans une certaine mesure. De toute évidence, le répondant SIPP déclare la date de réception du paiement et n'est pas au courant de la date d'émission tandis que l'inverse se produit dans le cas des dossiers relatifs au programme. Lorsque la date d'émission du chèque se situe vers la fin d'un mois, il peut être difficile de distinguer une erreur de télescopage croissant d'un écart légitime entre le mois de l'émission et le mois du paiement. Lorsqu'il y a des différences dans les définitions, par exemple pour cette question des dates d'émission/paiement, nous tenterons d'en présenter des modèles explicites dans nos analyses.

4. APPARIEMENT

4.1 Introduction

La qualité de l'appariement a une incidence considérable sur certaines des estimations des erreurs de réponse les plus importantes telles que la variance de l'erreur de réponse. Idéalement, les variables utilisées pour appairer les données de l'enquête et celles des dossiers seraient mesurées sans erreur et permettraient d'identifier un particulier. Il est entendu qu'un tel idéal n'est jamais atteint.

Toutefois, les variables dont nous disposons pour l'appariement des données de l'enquête et des dossiers devraient contribuer à réduire considérablement les erreurs d'appariement. Certaines d'entre elles, par exemple le numéro de sécurité sociale (SSN), permettent d'identifier de façon unique un particulier, même si d'autres renseignements comme l'adresse sont partiellement, illisibles, effacés ou manquants. Pour des raisons qui n'ont pas de rapport direct avec la présente étude (mais qui peuvent certainement lui être profitables), le Census Bureau a pris des mesures spéciales pour s'assurer que le SSN déclaré lors de la SIPP est complet et valide. On a vérifié les déclarations de toutes les personnes échantillonnées des vagues 1 et 2 qui ont fourni un SSN ou ont dit ne pas avoir de SSN et, au besoin, ces déclarations ont été corrigées par la Social Security Administration (administration de la sécurité sociale). À la suite de cette opération, Sater (1986) calcule que les données sur le SSN contenues dans le fichier SIPP sont valides pour environ 95 pour cent des personnes de l'échantillon SIPP qui en ont effectivement un.

La profusion d'autres données, nom de famille, prénom, numéro de voirie, nom de rue, nom de l'immeuble d'appartements, ville, code postal, sexe et date de naissance, suffit à assurer un appariement de grande qualité, même en l'absence d'un code d'identification unique tel que le SSN. En outre, pour nous aider à évaluer l'incidence de toute autre erreur d'appariement, le responsable de l'appariement du Census Bureau produit une mesure ordinaire de la valeur de l'appariement ou du non-appariement de chaque observation de l'enquête et de la donnée correspondante des dossiers administratifs.

L'étude de vérification des enregistrements porte sur un sous-ensemble de données SIPP fournies par le panel de 1984. Premièrement, l'échantillon est limité aux ménages résidant dans quatre Etats: la Floride, New York, la Pennsylvanie et le Wisconsin. Pour le panel de 1984, cela équivaut à approximativement 5,000 ménages. Deuxièmement, la période de référence de l'étude correspond seulement aux mois d'enquête des deux premières vagues du de 1984. La figure 1 montre la vague, le groupe de renouvellement, le mois d'interview et la période de référence qui correspondent aux données cibles de l'enquête.

Troisièmement, l'étude de vérification des enregistrements met l'accent sur la qualité des données sur la participation aux programmes et sur les sommes versées dans le cadre de certains programmes de transfert gouvernementaux. Elle compare les documents d'enquête et les dossiers administratifs concernant cinq programmes fédéraux (Federal Civil Service Retirement/Fonds de retraite des fonctionnaires fédéraux; Pell Grants/subventions Pell; Social Security — OASDI/sécurité de la vieillesse, pensions de survivant et assurance-invalidité; Supplemental Security Income/revenu supplémentaire de sécurité sociale; et Veterans' Compensation and Pensions/indemnités et pensions des anciens combattants) et quatre programmes administrés par les Etats (Aid to Families with Dependant Children/aide aux familles à faible revenu avec enfants à charge; les bons alimentaires; les prestations d'assurance-chômage et l'indemnisation des accidents du travail).

Nous avons limité l'étude à quatre Etats, soit la Floride, New York, la Pennsylvanie et le Wisconsin, afin d'avoir des tailles plus faciles à traiter. La sélection de ces Etats a été faite en fonction des critères suivants:

1. l'existence d'un système de dossiers automatisé, accessible et complet pour tous les programmes visés;
2. un vaste échantillon SIPP;
3. une diversité géographique appropriée, et
4. le consentement au partage des données individuelles pour les besoins de notre étude.

Ainsi, les Etats ont été choisis intentionnellement, aucune tentative n'a été faite pour choisir des Etats représentatifs du pays.

Nous avons demandé à chaque organisme participant de ces Etats de fournir des données d'identification sur toutes les personnes qui ont reçu un revenu au titre du programme cible entre mai 1983 et juin 1984 ainsi que les données sur les sommes reçues. La même demande a été adressée aux organismes fédéraux participants, mais seules les données sur les bénéficiaires résidant dans un des quatre Etats choisis devaient être fournies.

Vague	Groupe de renouvellement	Mois de la période de référence											
		Mois de l'interview	Juin	Juil.	Août	Sept.	Oct.	Nov.	Déc.	Janv.	Fév.	Mars	Avr.
1	1	Oct. 83	X										
	2	Nov. 83		X									
	3	Déc. 83			X								
	4	Janv. 84				X							
2	1	Fév. 84					X						
	2	Mars 84						X					
	3	Avr. 84							X				
3	41	Mai 84								X			
											X		
												X	
													X

Figure 1: Structure de l'enquête pour les données visées par l'étude de vérification des enregistrements SIPP.

Techniquement, le groupe de renouvellement 4 du panel de la SIPP de 1984 n'a pas fait l'objet d'une interview de la 2^e vague. Les répondants ne pouvaient pas se rendre compte qu'ils avaient «manqué» une interview; ils ont été soumis à l'interview de la 3^e vague au moment où ils auraient dû être interviewés dans le cadre de la 2^e vague. Aux fins de l'étude, l'interview de la 3^e vague pour le groupe de renouvellement 4 est identique à l'interview de la 2^e vague pour tous les autres groupes de renouvellement et il est inclus dans l'étude de vérification des enregistrements SIPP afin qu'il y ait deux interviews pour chaque cas de l'échantillon. Toute mention de la 2^e vague dans le présent article comprend les interviews de la 3^e vague pour ces membres du panel.

Medicaid; sur les transferts privés comme les prestations de retraite, les pensions alimentaires et les paiements pour garde d'enfant(s), la propriété de biens qui génèrent des revenus tels que des intérêts, des dividendes, des loyers et des redevances, et diverses autres sources de revenu, par exemple les successions.

2.2 Plan de collecte des données SIPP

La SIPP a été menée pour la première fois en octobre 1983 et visait un échantillon d'environ 25,000 unités de logement (le «panel de 1984») choisies pour représenter la population américaine hors institution. En février 1985, l'enquête a été menée auprès d'un nouveau panel légèrement moins important. Des panels supplémentaires sont censés être introduits dans le champ de l'enquête en février de chaque année pour toute la durée de l'enquête. En raison de restrictions budgétaires, la taille de l'échantillon des nouveaux panels s'élève actuellement à environ 15,000 ménages.

Chaque ménage de l'échantillon est interviewé au cours d'une visite à domicile tous les quatre mois pendant deux ans et demi, ce qui équivaut à un total de huit interviews par ménage. La période de référence pour chaque interview correspond aux quatre mois qui précèdent le mois d'interview. Lors de chaque visite, chaque membre du ménage de 15 ans et plus doit fournir des renseignements le concernant. Les déclarations faites au nom d'un membre absent au moment de la visite sont permises. Les renseignements concernant les déclarations faites par un enquêteur substitut sont consignés et disponibles à des fins d'analyse.

Pour faciliter les opérations sur le terrain, chaque panel-échantillon est divisé en quatre sous-échantillons (groupes de renouvellement) qui ont à peu près la même taille et dont un est interviewé chaque mois. Ainsi, une «vague» (cycle) d'interview est complétée après une période de quatre mois pour chaque panel. Grâce à ce plan d'enquête, les activités sur le terrain et les tâches de dépouillement se déroulent plus régulièrement; par contre, chaque groupe de renouvellement se trouve à avoir une période de référence de quatre mois légèrement différente. À partir de la deuxième vague d'interview du panel de 1984, des réinterviews sont menées auprès d'un petit échantillon de ménages qui doivent répondre à une sous-série de questions (y compris sur leur participation aux programmes). Ces données servent à relever les erreurs de l'intervieweur, mais peuvent peut-être aider à estimer les réponses incohérentes.

3. PLAN DE VÉRIFICATION DES ENREGISTREMENTS

L'objet de la vérification des enregistrements est de fournir une évaluation d'une partie des données sur le revenu recueillies dans le cadre de la SIPP. Nous allons maintenant mettre l'accent sur les principales caractéristiques du plan de vérification des enregistrements et traiter des aspects suivants: les échantillons, les dossiers administratifs, la méthode d'appariement et l'analyse.

3.1 Échantillons de la vérification des enregistrements

La vérification des enregistrements de la SIPP est fondée sur un plan «complet» plutôt qu'unidirectionnel. En d'autres mots, les enregistrements nous permettent de valider toutes les données observées de l'enquête. D'autres plans dont nous n'avons pas tenu compte consistaient :

1. à vérifier uniquement les enregistrements des personnes qui déclarent participer à un programme, ou
2. à tirer un échantillon de bénéficiaires connus et à interviewer ces derniers afin de déterminer si les renseignements qu'ils ont fournis sont vrais.

Ces deux types de plan sont incomplets et auraient faussé les estimations des paramètres des erreurs de réponse.

la date à laquelle ils ont reçu le paiement. De telles différences peuvent nuire considérablement à nos estimations chronologiques concernant, par exemple, les erreurs de réponse de téléscopage.

1.3.5 Absence d'expériences et de réentrevues

Les études de vérification d'enregistrements peuvent déceler les erreurs, mais n'indiquent pas comment les corriger. Pour déterminer l'efficacité d'un plan de sondage différent, il faut habituellement tester les autres plans qui pourraient être utilisés ou estimer les paramètres d'un modèle sous-jacent à partir duquel des plans de sondage peuvent être établis (par ex., un modèle des effets du manque de mémoire). Ainsi, un plan de vérification d'enregistrements peut servir à estimer et à comparer les erreurs de réponse chez les répondants eux-mêmes et chez les enquêtés-sustituts. Toutefois, sans hypothèses solides, un tel plan ne permet pas de déterminer dans quelle mesure les paramètres des erreurs de mesure changeraient si les règles de déclaration de l'enquête étaient modifiées (par ex., seuls les répondants visés pourraient fournir les réponses).

De même, une vérification des enregistrements faite sans réentrevue ou autre série de mesures indépendantes ne peut servir qu'à estimer un nombre limité de paramètres des erreurs de base. Ainsi, nos définitions initiales comportaient trois paramètres: la valeur réelle, les erreurs dans les données d'enquête et les erreurs dans les dossiers. Sans réentrevue (ou toute autre mesure indépendante), il n'y a que deux mesures qui permettent d'estimer ces trois inconnues. Une mesure supplémentaire comme la réentrevue peut aider à reconnaître les estimations des paramètres du modèle.

2. CARACTÉRISTIQUES DE L'ENQUÊTE SIPP

Nous ferons ici une brève description des principales caractéristiques de la SIPP, c'est-à-dire l'enquête sur le revenu et la participation aux programmes ou Survey of Income and Program Participation, avant de discuter du plan de vérification des enregistrements.

2.1 Aperçu du contenu de la SIPP

La SIPP a pour objet de fournir des renseignements plus précis sur la situation économique des particuliers et des ménages aux États-Unis. Elle permet de recueillir des données longitudinales complètes sur le revenu en espèces et toute forme d'aide autre que financière, sur l'admissibilité et la participation aux programmes de transfert du gouvernement, sur les éléments d'actif et de passif, sur l'activité et sur une foule de sujets connexes. Les données SIPP aident à l'évaluation du coût et de l'efficacité des programmes actuels de l'administration fédérale, de l'incidence éventuelle des changements que l'on propose d'apporter aux programmes et des effets réels de ces changements une fois qu'ils auront été appliqués. En général, le Census Bureau et d'autres organismes gouvernementaux qui ont encouragé et appuyé l'élaboration de la SIPP s'attendent à ce que celle-ci soit d'une aide précieuse pour la planification de la politique intérieure (Nelson et coll. 1985).

Les questions de base de la SIPP, qui sont répétées à chaque vague d'interview, portent sur l'activité sur le marché du travail ainsi que sur le revenu selon la source et le montant, y compris les paiements de transfert et l'aide autre que financière accordée dans le cadre de divers programmes pour chaque mois de la période de référence. Les questions de base portent sur presque cinquante sources de revenu, y compris les paiements de transfert gouvernementaux des fonds de retraite, les allocations d'invalidité et les prestations d'assurance-chômage et les programmes sociaux qui assurent, par exemple, une aide financière aux familles à faible revenu avec enfants à charge. Des renseignements sont aussi recueillis sur des programmes d'assistance non financière tels que le programme de distribution de bons alimentaires, Medicaid et

porter l'analyste à considérer jusqu'à la moitié de la variance des erreurs de réponse comme une erreur systématique de réponse et déterminer à l'avance la valeur positive ou négative de l'erreur systématique estimative de réponse si la variable mesurée est binaire (Marquis 1978). Il est nécessaire (mais non suffisant) que les plans soient complets pour obtenir des estimations non faussées des erreurs de réponse.

1.3.2. Erreurs d'appariement

L'objet de la vérification d'enregistrements consiste à appairer une à une les données d'enquête et des données de source administrative. C'est une opération qui est difficile à faire correctement, et les erreurs d'appariement (faux appariements, faux non-appariements) peuvent entraîner un biais dans les estimations des erreurs de mesure. Neter et ses collaborateurs (1965) ont montré que lorsqu'il n'y a pas de cas de non-appariement, les mauvais appariements donnent lieu à une erreur systématique par excès affectant la variance des erreurs de réponse. Pour ce qui est de la sûreté d'une mesure dichotomique (qui est une fonction de la variance des erreurs de réponse), l'estimation est réduite dans une proportion équivalente au taux des erreurs d'appariement (Marquis et coll. 1986). Par conséquent, il serait bon qu'il n'y ait qu'un minimum d'erreurs d'appariement et que l'on soit renseigné sur les erreurs qui restent.

1.3.3. Erreurs dans les dossiers administratifs

Comme nous l'avons mentionné précédemment, on peut habituellement compter sur le fait que les dossiers qui font l'objet d'une étude de vérification sont d'excellentes mesures du sujet d'intérêt. Si les hypothèses implicites concernant le biais relatif à la mesure fondée sur des dossiers et la variance des erreurs de la mesure se montrent inexacts, cela peut introduire un biais dans les estimations des erreurs de réponse. Par exemple, un biais touchant les données observées dans un dossier peut se manifester par un biais affectant les données observées de l'enquête, mais le biais serait de valeur contraire. Feather (1972) a décrit cet effet dans le cadre d'une vérification des dossiers des visites des médecins de la Saskatchewan. Le taux apparemment élevé de déclaration par excès enregistré au cours de l'enquête a été attribué au fait que le dossier contenait de l'information portant sur l'ensemble du traitement plutôt que sur les visites individuelles faites pour obtenir un diagnostic. De même, la présence d'une variance des erreurs de mesure dans les dossiers peut gonfler les estimations de la variance des erreurs de réponse dans l'enquête (Marquis 1978).

1.3.4 Différences dans les valeurs réelles

Des problèmes surviennent lorsque les définitions de l'enquête et du système de dossiers administratifs diffèrent. C'est souvent le cas lorsque des «comparaisons agrégatives» des estimations des paramètres de la population sont faites séparément par chaque source. Une des différences entre les deux systèmes est l'étendue des populations visées. Ainsi, la base de l'enquête peut être limitée à la population civile hors institution alors que les données des dossiers administratifs peuvent porter sur l'ensemble de la population. L'appariement de chaque cas peut contribuer à réduire les problèmes causés par la différence de champ d'observation, mais même les estimations faites à partir de ces études peuvent encore être affectées par le fait que les concepts ou les caractéristiques d'un concept diffèrent. Par exemple, Cox et Iachan (1987) font état des résultats d'une étude de comparaison des données sur l'état de santé déclarées par les répondants d'une enquête et celles des dossiers médicaux. Les auteurs concluent que le manque de concordance entre les résultats de l'enquête et les données des dossiers est attribuable en grande partie à des différences conceptuelles; en effet, l'enquête visait à recueillir des données au sujet de plaintes ayant donné lieu à une visite chez le médecin tandis que les dossiers médicaux portaient principalement sur le diagnostic final. Selon un exemple tiré de notre étude, les dossiers administratifs indiquent souvent la date d'émission d'un chèque établi en vue d'un paiement de transfert tandis que les répondants de notre enquête nous donnent

3. les interviews sans cadre de référence dans le temps donnent lieu à des déclarations excessives alors que ce problème ne se pose pas dans le cas des interviews qui en comportent;
4. la qualité des déclarations diminue lorsque l'interview se prolonge ou que le répondant fait partie de l'échantillon depuis un certain temps;
5. les gens sont fondamentalement paresseux et ont l'esprit tortueux, ils mentiront pour éviter d'avoir à répondre à une série détaillée de questions; et
6. les déclarations faites par le répondant visé sont préférables à celles qui sont faites par un enquêteur-substitut.

De fait, ces hypothèses sont devenues les fondements classiques des plans de sondage dans le monde occidental. Et pourtant, il est difficile d'étayer ces hypothèses à partir de vérifications d'enregistrements appropriées. Les expériences et autres techniques du genre constituent d'excellents moyens pour déterminer exactement les sources de la variation et pour démêler les problèmes d'estimation de la colinéarité, mais s'avèrent souvent inutiles et rarement suffisants pour évaluer un processus de mesure établi.

En somme, ces autres méthodes d'évaluation nous obligent à nous reposer en grande partie sur la supposition:

1. que la mesure initiale et la mesure d'évaluation sont indépendantes alors qu'elles sont clairement dépendantes;
2. qu'il existe un rapport entre la mesure initiale et un critère donné alors qu'il n'y a aucun lien externe objectif; et (ou)

3. que les processus cognitifs sont valables alors qu'ils ne sont pas étayés par des recherches. Pour les vérifications d'enregistrements, on utilise aussi des hypothèses pour évaluer les mesures. Ainsi, la façon habituelle d'estimer l'erreur systématique de réponse consiste à supposer qu'il n'y a pas de biais relatif aux dossiers ($\bar{u} = 0$) et à simplement calculer la moyenne des écarts entre les valeurs observées de l'enquête et du dossier qui sont comparées:

$$\text{déviati on systématique estimative de l'enquête} = \sum (S_i - R_i) / N.$$

Bien qu'il ne soit pas possible de corroborer directement l'hypothèse selon laquelle il n'y a pas de biais relatif aux dossiers, on peut effectuer des tests de sensibilité pour déterminer les effets sur les conclusions de l'évaluation de la non-vérification de cette hypothèse.

1.3 Points à retenir pour l'élaboration d'un plan de vérification d'enregistrements

Plusieurs points doivent être pris en considération lors de l'élaboration d'un plan de vérification d'enregistrements pour évaluer la qualité des données d'enquête, notamment les plans d'observation incomplets, les erreurs d'appariement, les erreurs dans les dossiers, les différences dans les valeurs réelles et l'absence de mesures répétitives ou de caractéristiques du plan d'expérience.

1.3.1. Plans d'observation incomplets

Les vérifications d'enregistrements antérieures ont souvent été faites à l'aide d'un plan unidirectionnel ou partiel établi pour la collecte des données, par exemple lorsque nous faisons enquête auprès des gens pour savoir s'ils possèdent une carte de bibliothèque et que nous vérifions les dossiers pour trouver ceux qui ont dit en avoir une, ou encore lorsque nous prélevons un échantillon à partir d'une liste de personnes souffrant d'une maladie chronique diagnostiquée et que nous les incluons dans une enquête afin de voir si ces personnes déclareront leur maladie dans le questionnaire d'enquête. Etant donné que ces plans partiels ne tiennent pas compte, dans les bonnes proportions, de la gamme complète des erreurs de réponse, ils produisent des estimations biaisées des paramètres classiques des erreurs de mesure, tels que l'erreur systématique de réponse et la variance des erreurs de réponse. Les plans unidirectionnels peuvent ne pas déceler une partie ou la totalité de la déviation systématique réelle de l'enquête,

dues à l'échantillonnage et d'erreurs d'échantillonnage. Nous ne présenterons pas un exposé technique, mais il est quand même nécessaire de définir en premier lieu certains des principaux termes que nous employons. Nous supposons que l'observation d'enquête à partir de l'unité d'échantillonnage i peut être exprimée comme étant la somme de la valeur réelle et d'une erreur, e_i :

$$\text{observation d'enquête}_i = \text{valeur réelle}_i + e_i.$$

Le biais moyen dans un ensemble d'observations d'enquête N , que nous appelons l'erreur systématique de réponse ou déviation systématique de l'enquête, est:

$$\bar{e} = \sum e_i / N,$$

et la variance des erreurs de réponse est simplement $\text{Var } e$. De même, le modèle servant à mesurer une observation fondée sur un dossier administratif est:

$$\text{dossier}_i = \text{valeur réelle}_i + u_i,$$

de sorte que le biais relatif au dossier est u et la variance des erreurs dans le dossier est $\text{Var } u$.

1.2 Comparaison des méthodes d'évaluation

La vérification des enregistrements présente des aspects qui peuvent être comparés à ceux d'autres méthodes d'évaluation telles que les réinterviews et les expériences. Les réinterviews et autres plans de mesures répétitives servent à estimer un ensemble très limité de paramètres des erreurs de mesure; c'est ce qu'on appelle habituellement la variance de réponse simple ou la variance des erreurs de réponse. Ces méthodes sont fondées implicitement sur des hypothèses fortes concernant les changements réels qui surviennent avec le temps et la valeur réelle ou le paramètre de biais (Marquis 1986).

Une solution que l'on utilise souvent consiste à introduire des mesures de la valeur réelle dans le programme de réinterview, par exemple en comparant des réponses divergentes obtenues avec celles que fournit un répondant bien informé ou en posant des questions beaucoup plus détaillées et précises au cours de la réinterview. Mais la validité de telles mesures peut être mise en doute. Bailier (1986) et Koons (1973) ont démontré, par exemple, que les réponses de réinterview ayant fait l'objet d'un rapprochement sont affectées d'une erreur systématique. Et bien que l'on préfère souvent un questionnaire comportant des questions précises et détaillées à une approche plus globale, il n'existe aucune preuve en soi indiquant qu'une telle méthode permettrait de supprimer, ou même de réduire, le biais. Les vérifications d'enregistrements peuvent fournir de meilleures données-critères, nécessitant des hypothèses beaucoup moins solides (et peut-être plus réalistes) pour l'estimation de la qualité des données d'enquête.

L'expérience est une autre méthode d'évaluation des aspects d'une enquête. On peut avoir recours, par exemple, à un plan factoriel entièrement croisé ou à un plan de sondage à réseaux superposés pour l'affectation des intervieweurs. Les analystes comparent les groupes visés par l'expérience et plus précisément, les statistiques concernant, par exemple, les moyennes ou les proportions relatives aux sujets spécialisés, et déterminent les taux de déclaration relatifs aux domaines spécialisés d'intérêt qui sont obtenus par les différentes méthodes. Toutefois, il y a controverse lorsqu'il faut établir laquelle des méthodes est la meilleure pour effectuer des mesures. Cette difficulté se trouve considérablement aplaniée lorsque nous disposons de données-critères, par exemple les dossiers administratifs.

Sans données servant de critères, l'analyste doit souvent s'en remettre à des hypothèses reconnues au sujet des erreurs de mesure, par exemple:

1. plus il y a de renseignements demandés, meilleurs sont ces renseignements;
2. la possibilité d'oublier des faits importants augmente avec le temps;

Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes

JEFFREY C. MOORE et KENT H. MARQUIS¹

RÉSUMÉ

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une importante nouvelle enquête par panel du Census Bureau destinée à fournir des données sur la situation économique des particuliers et des familles aux États-Unis. La donnée de base de la SIPP est le revenu mensuel qui est déclaré pour chacun des quatre mois de la période de référence qui précède le mois de l'interview. L'étude de vérification des enregistrements SIPP utilise les données des dossiers administratifs pour évaluer la qualité des estimations de la SIPP pour un grand nombre de sources de revenus et de programmes de transfert. Le projet utilise des techniques informatiques d'appariement des enregistrements pour identifier les personnes échantillonnées dans quatre États qui, selon les dossiers, ont reçu des paiements de n'importe lequel de neuf programmes administrés par l'État ou par le gouvernement fédéral. On compare ensuite les dates et les montants des paiements déclarés lors de l'enquête aux données correspondantes des dossiers officiels. Le document décrit le projet en détail et présente quelques-unes des premières conclusions.

MOTS CLÉS: SIPP; vérification des enregistrements; couplage des enregistrements; validité des réponses par enquêtes.

1. INTRODUCTION

La présente communication traite des questions relatives à l'utilisation des données de dossiers administratifs pour évaluer la qualité des estimations fondées sur des données d'enquête et décrit une application précise aux données de la SIPP aux États-Unis. L'appariement des données des dossiers administratifs et des observations de l'enquête effectuée pour chaque cas et que nous appelons «vérification des enregistrements», fournit des renseignements utiles aux personnes chargées de l'élaboration de l'enquête et à celles qui utilisent ses données. Une vérification des enregistrements permet à l'analyste de faire une gamme complète d'estimations des paramètres des erreurs de mesure à des fins d'évaluation. Ces estimations facilitent à leur tour la réalisation de deux genres d'activités fondamentales:

1. la quantification des effets des erreurs de mesure sur les estimations concernant des sujets spécialisés tels que les moyennes, les proportions, les coefficients de corrélation et les coefficients de régression à plusieurs variables (et peut-être l'ajustement des estimations afin de corriger les erreurs de mesure), et
2. l'établissement de plans de sondage plus efficaces qui tiennent compte, par exemple, de l'équilibre à maintenir entre la qualité des mesures et les coûts.

1.1 Principaux termes utilisés

Nous nous concentrons ici sur les erreurs de mesure (ou de «réponse»), bien que la méthode de vérification des enregistrements puisse aussi servir à évaluer d'autres types d'erreurs non

¹ Jeffrey C. Moore et Kent H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, pièce 433, Washington Plaza Building, Washington, DC 20233. Le présent article est la version révisée d'une communication présentée au Symposium international sur les utilisations statistiques des données administratives les 23-25 novembre 1987. Cette communication donne le point de vue des auteurs et ne reflète pas nécessairement la politique officielle du Census Bureau ou les opinions de la direction.

les programmeurs et les ressources budgétaires limitées dont il dispose pour effectuer la vérification. De plus, le recours aux micro-ordinateurs, parce qu'il offre beaucoup de souplesse d'utilisation, a permis d'accroître le rôle des analystes et de diminuer la charge de travail des programmeurs. Les échanges entre le Censur Bureau et l'IRS étant nombreux, les personnes concernées sont plus sensibles à l'importance de la qualité des données. Ces échanges permettent également de mettre au point des procédures efficaces de transmission des données. Enfin, grâce à cet effort collectif pour régler les problèmes, le Censur Bureau est maintenant assuré de recevoir les données requises pour mener à bien les recensements économiques et agricole de 1987.

REMERCIEMENTS

Nous voulons remercier les arbitres pour leurs commentaires et suggestions.

Tableau 5

Modes de déclaration des éléments d'information des formules I120S, 1986 (chiffres pondérés)

Éléments d'information	Pourcentage de formules I120S		Objectif
	reçu	exigé	
EIN	Espaces laissés en blanc, zéros, données non numériques	0.0	Moins de 1.0
	Code IRD invalide	0.0	Moins de 1.0
	CODE PBA	0.0	Moins de 6.0
	Espaces laissés en blanc ou données non numériques	0.0	Moins de 6.0
REVENUS OU CHIFFRE DES VENTES BRUTS	Espaces laissés en blanc, données non numériques, codes non classés ou codes PBA invalides	11.5	Moins de 18.0
	MOINS RENDUS ET RABAIS		
	Espaces laissés en blanc, zéros ou données non numériques	20.9	Moins de 40.0
	Enregistrements pour lesquels un chiffre est fourni, pourcentage dans chacune des tranches ci-après:		
EXERCICE FINANCIER	- moins de \$100,000	45.7	30.0 — 60.0
	- \$100,000 ou plus mais moins de \$500,000	36.9	20.0 — 50.0
	- \$500,000 ou plus	17.4	10.0 — 30.0
	Espaces laissés en blanc, zéros ou données non numériques	0.0	Moins de 1.0

L'utilisation combinée de micro-ordinateurs et d'un gros ordinateur a permis au Census Bureau de vérifier, d'une manière efficace et exhaustive, les très gros fichiers de données de l'IRS et d'en assurer la qualité. En outre, le système a allégé la charge des programmeurs et il continuera de le faire parce qu'une bonne partie du travail peut être faite sur micros-ordinateurs, par des non-spécialistes. Par ailleurs, puisque le système a permis de réduire l'utilisation du gros ordinateur et la charge de travail des programmeurs, ces derniers peuvent concentrer leurs efforts sur les totalisations de base. Enfin, en ce qui concerne l'analyse des données, le système permet que la révision puisse être effectuée par du personnel de différents niveaux. Les états récapitulatifs permettent aux gestionnaires de déterminer rapidement et efficacement si les délais de transmission sont respectés et si la qualité des données est satisfaisante. Les analystes, pour leur part, consultent les rapports hebdomadaires plus détaillés qui leur indiquent, le cas échéant, les ensembles de données nécessitant un examen plus approfondi.

5. RÉSUMÉ

Le Census Bureau a mis au point un système global de contrôle de la qualité qui permet de détecter les problèmes risquant de nuire à la qualité des données. Avec le système, il est possible de traiter de très gros fichiers de données de l'IRS. En outre, comme le système favorise la collaboration, il permet au Census Bureau de régler rapidement les problèmes de manière à assurer la transmission adéquate des données. Les critères de qualité et les délais de transmission sont définis conjointement par le Bureau et l'IRS. Le respect de ces critères est vérifié par l'ordinateur. Par ailleurs, grâce à l'informatisation, le Bureau est en mesure d'utiliser au mieux

Tableau 4a			
Répartition pondérée des Formules 1120 (1986), en nombre, selon la date			
Date	Formules 1120		Objectif non atteint
	Nombre reçu	Nombre exigé	
Fin mars 1987	326,500	303,000	
Fin avril 1987	697,600	760,000	
Fin mai 1987		988,000	
Fin juin 1987		1,190,000	
Fin juillet 1987		1,418,000	
Fin août 1987		1,621,000	
Fin janvier 1988		2,077,000	
Fin octobre 1988		2,533,000	

Tableau 4b			
Répartition pondérée des Formules 1120S (1986), en nombre, selon la date			
Date	Formules 1120S		Objectif non atteint
	Nombre reçu	Nombre exigé	
Fin mars 1987	103,350	90,000	
Fin avril 1987	328,850	225,000	
Fin mai 1987		292,000	
Fin juin 1987		352,000	
Fin juillet 1987		420,000	
Fin août 1987		480,000	
Fin janvier 1988		615,000	
Fin octobre 1988		750,000	

entraîner des frais de traitement informatique supplémentaires. Aux recensements économiques et agricole, le Bureau a relevé un cas semblable (voir tableaux 4a et 4b): à la fin de mai 1987, le Census Bureau avait reçu environ 697,600 Formules 1120 (sociétés); le nombre requis étant de 760,000. L'objectif n'a donc pas été atteint. Par contre, le Bureau a reçu un nombre beaucoup plus élevé de Formules 1120S («S Corporations») que le nombre exigé (environ 328,850 au lieu de 225,000). La cause de ce changement est la suivante: avec l'adoption de la nouvelle loi fiscale, il devenait plus avantageux pour les entreprises de remplir une Formule 1120S qu'une Formule 1120. Ainsi, bien qu'il n'y ait eu aucune erreur dans les données, il était important que le Bureau relève ce changement pour pouvoir modifier ses méthodes de traitement. Nous sommes en train de mettre au point des mesures pour tenir compte de cet accroissement du nombre des «S Corporations». Le tableau 5 est un exemple des divers tableaux que nous utilisons pour étudier la qualité des données. Comme on peut le voir, la qualité des données répond aux exigences. Si, pour un élément d'information donné, nos critères n'étaient pas respectés, un indicateur serait affiché pour qu'un analyste fasse les recherches nécessaires.

Le système informatisé de contrôle de la qualité sera tout à fait prêt pour effectuer le traitement des fichiers de l'IRS de 1987. Des systèmes pilotes ont été utilisés (et continuent de l'être) pour évaluer les fichiers de 1985 et de 1986. C'est grâce à ces systèmes que le Bureau a pu, en vue des recensements agricole et économique, mener à bien la vérification des données administratives de 1985 et de 1986 transmises par l'IRS.

Outre cet ensemble complet de tableaux cumulatifs, nous produisons également un ensemble de tableaux de résultats. Ces tableaux contiennent des comparaisons détaillées de certaines données clés pour le cycle en cours. Le tableau 3 est un exemple des nombreux tableaux des résultats produits par le système. On y trouve le nombre et le pourcentage pondérés de formules 1040 — annexe F par centre de traitement de même que le pourcentage attendu de formules. Comme on peut le voir, les chiffres cumulatifs sont raisonnables et répondent aux critères établis. Si des incohérences avaient été relevées, un indicateur aurait été affiché en regard du nom du centre de traitement concerné. Le dernier type de rapport produit par le système automatisé est le rapport de la situation qui est un bilan détaillé du nombre cumulatif de fichiers envoyés par l'IRS. Dans ce rapport, on compare la qualité globale des ensembles de données aux critères établis (délais pour la transmission des données et exigences qualitatives). Ces rapports sont transmis à l'IRS une fois par mois environ. Comme nous l'avons souligné, ces rapports, qui constituent un état récapitulatif de la qualité des données, favorisent les échanges entre le personnel du Bureau et celui de l'IRS.

4. RÉSULTATS DU CONTRÔLE DE LA QUALITÉ

Les rapports de la situation permettent au Census Bureau et à l'IRS de déceler les problèmes au fur et à mesure et ils facilitent la collaboration. Il est à noter cependant que lorsque nos critères ne sont pas respectés (délais ou qualité), la plupart du temps c'est en raison d'un changement dans les modes de déclaration des entreprises. Des cas semblables ne posent aucune difficulté pour l'IRS (il n'a pas de mesure corrective à prendre) mais, pour le Bureau, ils peuvent

Tableau 3

Répartition pondérée des Formules 1040, annexe F (1986), en pourcentage, selon le centre de traitement

Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	2,087,200	176,700	71,600	374,900	262,100	358,600
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							
1986	En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5
	Objectif visé	100.0	8.5	3.0	18.5	11.5	17.5
Objectif non atteint ¹							
Année	d'imposition	Annexes	Centre de traitement				
			F reçues	Atlanta	Philadelphie	Austin	Cincinnati
1986	En nombre	118,800	343,200	40,300	288,100	52,500	400
	En Pourcentage	100.0	8.5	3.4	18.0	12.6	17.2
Objectif visé							

données administratives. Cette procédure a permis au Censur Bureau de concevoir un programme pour micro-ordinateur permettant de construire des images de tableaux pour n'importe quel genre de fichier de données administratives. Une fois achevées, les images de tableaux sont transférées au gros ordinateur et sont utilisées par les programmeurs pour aligner les données contenues dans les fichiers. Le travail de programmation qu'exigent les tableaux de contrôle de la qualité est de beaucoup simplifié du fait que la construction des images de tableaux est effectuée par des non-spécialistes. Les programmeurs travaillant sur gros ordinateur peuvent donc concentrer leurs efforts uniquement sur l'établissement des totalisations. Le tableau 2 est un exemple des divers tableaux produits sur gros ordinateur pour chaque type d'entrepris. Ce tableau fait voir la répartition pondérée des formules 1040, annexe C, selon le centre de traitement et la tranche de revenu net.

Le gros ordinateur n'effectue que les totalisations de base (mise à jour des totalisations). Ces dernières sont ensuite transférées sur des micro-ordinateurs où toutes les autres opérations de vérification sont effectuées. Parmi celles-ci, mentionnons le calcul des pourcentages dans les nouvelles totalisations, la production de totalisations cumulatives, la vérification de données clés et la production de rapports sur l'état d'avancement des travaux de vérification. Cette méthode de traitement systématique, qui s'appuie essentiellement sur l'utilisation de micro-ordinateurs, offre plus de souplesse pour l'examen des données et permet de réduire la charge de travail des programmeurs.

Les totalisations produites sur gros ordinateur sont transférées dans un tableau préformaté exploité sur micro-ordinateur. Ce tableau contient également les critères définis par le Censur Bureau relativement aux ensembles de données administratives. Le micro-ordinateur effectue automatiquement la vérification des données du tableau. Les écarts entre les données et les exigences établies sont indiqués dans les tableaux de résultats. Les deux principaux avantages du système de vérification sont les suivants:

1. Il nous permet de détecter aisément les problèmes. Les données qui ne satisfont pas aux normes sont signalées par un indicateur afin qu'un analyste les examine. On risque donc moins d'oublier des erreurs.

2. Il attire notre attention sur des données qu'il faudrait examiner de plus près. Souvent, bien que toutes nos exigences aient été remplies, les tableaux de résultats nous mettent sur la piste de certains problèmes (par exemple, le système nous permet d'examiner plus en détail certaines tendances inattendues). En fait, les tableaux de résultats nous permettent de concentrer notre attention sur les données qui peuvent poser des difficultés. Il peut s'agir de demander à un centre de traitement de l'IRS d'examiner à nouveau les données ou encore, de transférer du gros ordinateur sur micro les enregistrements présents tant certaines caractéristiques. Ensuite, nous examinons ces enregistrements à la main afin de cerner les problèmes.

Comme nous l'avons souligné, nous avons défini les exigences qualitatives de base auxquelles doivent satisfaire les données pour les recensements économique et agricole de 1987. Notre système automatisé (comparaison des données fournies aux critères établis), nous permet de déterminer immédiatement si les données sont de qualité acceptable.

Après examen et vérification des données du cycle en cours, des tableaux cumulatifs sont produits sur micro-ordinateur. Le fait d'avoir recours à des micros plutôt qu'à un gros ordinateur pour la production de ces tableaux permet une utilisation plus efficace de nos ressources. Premièrement, le système nous évite d'avoir à garder en mémoire du gros ordinateur des fichiers cumulatifs, ce qui réduit les coûts d'ordinateur. Auparavant, ces fichiers cumulatifs étaient conservés dans le gros ordinateur et étaient combinés aux fichiers du cycle en cours pour produire de nouveaux tableaux cumulatifs. En utilisant les micro-ordinateurs, nous avons pu intégrer au tableau des formules simples qui nous permettent de créer des tableaux cumulatifs à très peu de frais. Deuxièmement, la production de ces tableaux cumulatifs n'engage pas la création de programmes pour gros ordinateur. Des imprimés des tableaux cumulatifs sont produits et conservés pour analyse ou référence.

Tableau 2

Répartition pondérée des formules 1040, annexe C selon le centre de traitement et la tranche de revenu net¹

Tranche de revenu net (000)		Centre de traitement					
		Total		< 0		en blanc	
				ou 0		1—	
Ensemble des centres	1,327,100	200	52,200	149,300	73,900	98,100	11,000
Atlanta	133,200	0	5,100	16,500	6,300	11,000	9,600
Philadelphie	132,100	100	4,200	11,300	5,300	12,900	8,500
Austin	147,600	0	6,300	20,900	9,900	12,900	9,800
Cincinnati	153,100	0	5,300	14,900	8,700	9,800	8,500
Kansas City	119,500	0	5,500	16,700	7,500	8,500	8,200
Andover	111,100	0	3,800	9,800	6,700	8,200	11,600
Ogden	162,300	0	7,500	20,200	7,900	11,600	10,000
Brookhaven	111,900	0	4,400	12,600	7,100	10,000	8,600
Memphis	136,500	100	4,700	14,700	6,700	7,900	7,900
Fresno	0	0	5,400	11,700	7,800	0	0
Autres	100	0	0	0	0	0	0

¹ Revenus bruts moins les rendus et rabais sur les ventes.

Tranche de revenu net (000)		Centre de traitement					
		10,000—		25,000—		50,000—	
		24,999		49,999		99,999	
Ensemble des centres	168,600	185,500	225,100	243,400	87,400	43,400	4,700
Atlanta	17,000	19,800	22,200	22,200	8,400	4,700	4,200
Philadelphie	17,800	19,800	22,700	27,000	10,100	4,200	4,400
Austin	18,700	18,500	22,000	24,900	9,100	4,400	5,800
Cincinnati	20,500	20,700	27,300	30,500	9,600	5,800	3,800
Kansas City	16,200	15,900	20,700	18,300	6,400	3,800	4,000
Andover	13,600	16,700	19,500	20,000	8,800	4,000	4,200
Ogden	17,800	19,500	28,800	33,600	11,200	4,200	3,300
Brookhaven	16,400	19,700	20,400	19,400	6,400	3,300	2,900
Memphis	15,100	14,700	18,600	19,000	6,800	2,900	6,100
Fresno	15,500	20,200	22,900	28,400	10,600	6,100	0
Autres	0	0	0	100	0	0	0

3. PARTICULARITES DU SYSTEME

Les fichiers de données que le Bureau reçoit chaque semaine sont d'abord soumis à un contrôle qualitatif sur gros ordinateur. Les programmes utilisés sont mis au point bien avant l'arrivée des fichiers et ils servent à produire les premiers tableaux de contrôle de la qualité, qui constituent la pierre angulaire de tout le système. Auparavant, les programmeurs travaillant sur gros ordinateur étaient chargés de créer les tableaux dans leur ensemble, ce qui comprend les cellules de données et le texte qui s'y rapporte (en-têtes et colonnes principales). Toutefois, dans les programmes de production de tableaux pour le recensement de 1987, ces deux composantes seront traitées séparément. La production des tableaux continuera de se faire sur gros ordinateur tandis que le texte accompagnant les tableaux sera produit à l'aide de micro-ordinateurs par des employés qui ne sont pas des spécialistes de la programmation. Une procédure a été mise au point pour permettre la production de tableaux pour tous les fichiers de

ont été faites sur gros ordinateur. L'utilisation de l'ordinateur central coûte cher, d'autant plus que les fichiers sont de plus en plus importants.

3. Manque de communication entre les organismes

Par le passé, le manque de communication entre les organismes intéressés a nui à la qualité des données. Pour corriger ce problème, il était indispensable de définir clairement les axes de communication devant exister entre le Bureau et l'organisme qui lui fournit les données, et cela, pour chacune des étapes du travail. Le Bureau doit d'abord définir les fichiers de données et les enregistrements dont il a besoin puis s'entendre avec l'organisme sur les données qui peuvent lui être fournies. Certaines des données demandées ne sont peut-être pas disponibles ou encore peuvent être trop coûteuses à produire. Les difficultés doivent être aplanies au fur et à mesure car des retards dans la transmission des données pourraient compromettre l'utilité. En outre, les organismes doivent s'entendre sur la qualité et la quantité des données désirées. Les exigences du Bureau touchant les données requises doivent être clairement quantifiées.

C'est pour apporter une solution globale au problème posé dans le passé par la qualité des données administratives que le Bureau a décidé d'élaborer et de mettre en oeuvre son système informatisé de contrôle de la qualité: le système permet de vérifier des fichiers importants et complexes de l'IRS, favorise les échanges et permet la détection immédiate des erreurs. La composition de «vérification informatisée» est la plus importante du système. En gros, le Bureau définit les critères ou exigences de base auxquels doivent répondre les données fournies par l'IRS. Il compare ensuite les modes de déclaration aux critères établis puis procède, sur micro-ordinateur, à la vérification systématique des données. Le Bureau produit ensuite un rapport de la situation dans lequel il indique si les données répondent ou non aux exigences.

Le personnel du Bureau définit ses exigences bien avant que les données ne lui soient transmises. L'IRS dispose donc de tout le temps voulu pour déterminer s'il peut raisonnablement fournir les données demandées et, au besoin, pour modifier les demandes. Les exigences définies par le Bureau portent sur les deux points suivants: fréquence de transmission et qualité des données. Pour ce qui est de la fréquence de transmission, le Bureau établit une estimation du nombre total de déclarations qu'il souhaite obtenir pour chaque type d'entreprise avec la date à laquelle il désire les recevoir. Pour ce qui est des exigences qualitatives, le Bureau définit les modes de déclaration désirés pour chaque type d'élément d'information.

Le système de vérification automatisé facilite l'analyse des données. Une série de tableaux de résultats sont produits afin de comparer les données fournies aux critères établis. Des indicateurs permettent de relever les données qui ne satisfont pas aux exigences. Cette façon de procéder réduit les risques d'omission pendant le processus d'analyse.

Chaque mois, le Bureau envoie à l'IRS des rapports de la situation dans lesquels les modes de déclaration des données sont comparés aux critères établis. Ces rapports constituent des sous-ensembles des tableaux détaillés des résultats et ne contiennent que les exigences de base pour les ensembles de données fournis par l'IRS. Ces rapports favorisent la communication, entre le Bureau et l'IRS. Les données posant des problèmes y sont indiquées et le Bureau et l'IRS doivent décider immédiatement des mesures correctives ou des mesures de récupération à prendre afin d'éviter de compromettre les résultats du recensement. À cet égard, il est indispensable que les mesures nécessaires soient prises dans les plus brefs délais parce que l'IRS ne conserve pas ses bandes de données indéfiniment. Si les erreurs ne sont pas détectées rapidement et si les mesures correctives ne sont pas prises à temps, la récupération risque d'être impossible ou encore d'être extrêmement coûteuse.

Le système de contrôle de la qualité ne nous permet pas de nous assurer que les données administratives ne poseront plus jamais de difficultés. Toutefois, il nous permet de définir très clairement nos exigences, de sorte que les caractéristiques des ensembles de données ne sont pas laissées au hasard. Il nous permet par ailleurs de déceler les erreurs le plus rapidement possible.

Tableau 1

Nombre approximatif d'enregistrements tirés de dossiers administratifs utilisés aux recensements économiques et agricole de 1987, selon le genre de formule et l'année d'imposition

Genre de fichier		1985	1986	1987
Nombre d'enregistrements				
Déclaration d'impôt sur le revenu des entreprises:				
Formule 1040, annexe C	—	11,750,000	12,500,000	30,881,000
Formule 1040, annexe F	2,450,000	—	2,450,000	—
Formule 1040, annexe SE	—	—	—	10,000,000
Formule 1120	42,000	2,550,000	2,650,000	210,000
Formule 1120-A	—	200,000	11,000	11,000
Formule 1120F	—	17,000	900,000	950,000
Formule 1065	108,000	1,750,000	1,800,000	1,800,000
Formule 990	—	380,000	400,000	400,000
Formule 990-PF	—	35,000	35,000	35,000
Formule 990-T	—	25,000	25,000	700,000
Formule 1120S, annexe K-1	—	—	—	1,600,000
Formule 1065, annexe K-1	—	—	—	45,050,000
Fichiers des données fiscales (données annuelles):				
IRS Fichier principal des entreprises	41,950,000	43,500,000	26,000,000	18,000,000
IRS Fichier de la paye et du personnel	17,000,000	17,500,000	1,050,000	1,050,000
SSA Fichier des nouvelles entreprises	950,000	1,000,000	—	—
Total	44,567,000	63,551,000	75,931,000	—

au total. On trouvera dans le tableau 1 le nombre approximatif d'enregistrements qui seront tirés, aux fins des recensements économiques et agricole de 1987, des diverses formules d'impôt. Il est clair que le nombre d'enregistrements que nous recevons durant une année de recensement est très élevé. Toutefois, la raison pour laquelle nous avons envisagé de mettre au point un système de contrôle de la qualité aussi perfectionné ne tient pas uniquement à la quantité des données à traiter mais aussi à la complexité du traitement. Un enregistrement contient souvent plusieurs données élémentaires et chacune d'elles vient accroître la complexité (niveau de détail) de l'enregistrement lui-même et de l'ensemble du fichier. En outre, les différentes formules d'impôt ne contiennent pas toutes les mêmes éléments d'information et les revenus n'y sont pas tous déclarés de la même façon. Par conséquent, non seulement faut-il vérifier la qualité de plus de 75 millions d'enregistrements mais aussi la qualité des données élémentaires qu'ils contiennent. Souignons enfin que les entreprises transmettent leurs déclarations d'impôt à l'un des dix centres de traitement de l'IRS. Chaque centre assure le traitement des déclarations qu'il reçoit et la qualité des données qui nous sont transmises peut varier. Le Bureau vérifie donc individuellement la qualité des données transmises par chacun d'eux.

2. Contraintes budgétaires

Les contraintes budgétaires sont un autre facteur important qui vient accroître la difficulté du contrôle de la qualité des données administratives. Compte tenu de la politique générale de restriction des dépenses du gouvernement, le Bureau essaie d'offrir des services plus nombreux à un moindre coût. La charge de travail en programmation est beaucoup plus lourde pendant les années de recensement, mais le nombre des programmeurs n'est pas augmenté proportionnellement. La vérification de la qualité, qui dépend dans une large mesure des diverses ressources existantes, peut en souffrir. Il faut également souligner que jusqu'à maintenant, la plupart des tâches de vérification de la qualité

programmation sont limitées et coûteuses, il serait difficile d'adopter un système exhaustif de contrôle exploité entièrement sur gros ordinateur. En outre, vu le caractère confidentiel des données et les conséquences possibles des erreurs, nous avons conclu qu'il fallait mettre au point un système de contrôle plus «sophistiqué». Le Census Bureau a réussi à élaborer un système complet pour gérer les différentes étapes du processus de vérification des données administratives. Ce système informatisé nous permettra d'effectuer une vérification plus approfondie des données malgré les contraintes budgétaires qui nous sont imposées et les ressources de programmation limitées dont nous disposons.

Le système intègre la technologie des gros ordinateurs à celle des micro-ordinateurs. Nous avons défini des critères ou exigences de base quant au genre de données administratives que nous désirons obtenir. Nous stockons ces critères dans la mémoire des micro-ordinateurs. Après exécution du programme de contrôle de la qualité sur gros ordinateur, nous faisons transférer les résultats sur les micro-ordinateurs. Les modes de déclaration des données sont ensuite comparés à nos exigences. La vérification proprement dite (examen des données, détection des erreurs, production de rapports) est faite sur micro-ordinateur. L'utilisation des micro-ordinateurs par opposition au recours exclusif à un gros ordinateur, permet d'alléger la charge des programmeurs, de faciliter le travail des analystes (plus de souplesse) et de réduire les coûts de traitement sur gros ordinateur. En outre, la composante «contrôle-vérification» du système assure une vérification rigoureuse et exhaustive des données. Enfin, bien que ce système ait été conçu spécialement pour la vérification des données provenant des déclarations d'impôt sur le revenu des entreprises aux fins du recensement économique, il peut être modifié (et il le sera après 1988) pour effectuer la vérification de toutes les données administratives que nous recevons.

2. APERÇU DU SYSTÈME DANS UNE PERSPECTIVE DE GESTION

Le Census Bureau a beaucoup recours aux dossiers administratifs pour produire des statistiques. En fait, l'importance de cette source de données n'a cessé de croître avec les années. Cet état de choses tient à la nécessité de produire des statistiques plus nombreuses et de meilleure qualité, de réduire au maximum le fardeau de réponse des entreprises privées et d'optimiser l'utilisation de nos ressources humaines et financières.

Ces dernières années, les données provenant des dossiers administratifs ont été, dans l'ensemble, jugées d'excellente qualité. Toutefois, les données tirées des déclarations d'impôt des entreprises qui nous ont été fournies en 1982 par l'IRS nous ont causé certains problèmes. Nous pensons, en particulier, à la qualité des principaux codes d'activité économique pour les entreprises à propriété unique. C'est en raison de ce problème, qu'au recensement de 1982, le Census Bureau n'a pu produire qu'une quantité limitée de statistiques sur les non-employeurs. Si nos programmes de contrôle de la qualité avaient été plus perfectionnés, nous aurions pu cerner plus rapidement les erreurs et en réduire les effets.

Lorsqu'est venu le moment de préparer le recensement de 1987, nous nous sommes rendu compte qu'il fallait prendre des mesures supplémentaires pour assurer la qualité des données administratives fournies par l'IRS. Ce dont nous avions besoin, c'était d'un système global de contrôle nous permettant de composer avec les trois principaux facteurs qui, dans le passé, ont compromis la qualité des ensembles de données administratives. Ces trois sources de problèmes sont les suivantes:

1. Volume élevé des données administratives

L'IRS nous fournira des enregistrements provenant des déclarations d'impôt produites par les entreprises pour 1987 (reçues en 1988). Ces enregistrements sont tirés des déclarations d'entreprises ayant différentes formes juridiques, notamment, des sociétés de capitaux, des sociétés dites «S Corporations» (petites sociétés), des sociétés étrangères, des sociétés de personnes, des entreprises à but non lucratif et des entreprises à propriété unique. En 1988, le Census Bureau prévoit recevoir au-delà de 75 millions d'enregistrements

Contrôle automatisé de la qualité des données provenant de dossiers administratifs

JAMES R. JONAS et PAUL S. HANCZARYK¹

RÉSUMÉ

Aux fins de son programme de la statistique économique, le Census Bureau fait grande utilisation des dossiers administratifs. Le nombre de dossiers traités chaque année est considérable, mais il l'est encore plus durant les années de recensement. Pour assurer la qualité de toutes les données qu'il reçoit, le Census Bureau a mis au point des programmes de contrôle sur gros ordinateur. Toutefois, étant donné le très grand nombre de tableaux de vérification requis, et compte tenu du fait que les ressources nécessaires à la programmation sont limitées et coûteuses, il serait difficile d'adopter un système exhaustif de contrôle exploitant entièrement sur gros ordinateur. En outre, vu le caractère confidentiel des données et les conséquences possibles des erreurs, le Census Bureau a conclu qu'il fallait mettre au point un système de contrôle plus «sophistiqué», qui s'appuie sur la technologie des micro-ordinateurs. La division des enquêtes économiques est en train de mettre au point un tel système qui sera utilisé pour traiter les fichiers de données administratives de 1987. Le système intègre la technologie des gros ordinateurs à celle des micro-ordinateurs. Les fichiers de données administratives qui arrivent chaque semaine au Bureau sont d'abord traités sur gros ordinateur. Les données ainsi traitées sont transférées sur micro-ordinateur et sont formatées pour être exploitées par un tableau. La vérification systématique de la qualité des données (examen de données, détection des erreurs, production de rapports) est faite sur micro-ordinateur, par le tableau. L'utilisation combinée d'un gros ordinateur et de micro-ordinateurs permet d'alléger la charge des programmeurs, de faciliter le travail des analystes (plus de souplesse) et de réduire les coûts du traitement sur gros ordinateur. Enfin, la composante «contrôle-vérification» du système assure une vérification rigoureuse et exhaustive des données.

MOTS CLÉS: Intégration de la technologie des micro-ordinateurs et des gros ordinateurs; vérification systématique des données; actualité des données.

1. INTRODUCTION

Aux fins de son programme de la statistique économique, le *Bureau of the Census* fait grande utilisation des données administratives, en particulier des données provenant des dossiers des déclarations d'impôt sur le revenu des entreprises du *Internal Revenue Service* (IRS) et, dans une moindre mesure, des dossiers de la *Social Security Administration* (SSA). Dans les années où ont lieu le recensement agricole et le recensement économique, le Census Bureau reçoit une quantité beaucoup plus considérable de données administratives. Ces données nous permettent de mener nos recensements économique et agricole de manière efficace et dans les meilleurs délais de même que de réduire au maximum le fardeau de réponse des entreprises et des exploitants agricoles. La réussite de nos programmes de la statistique économique et de la statistique agricole dépend, dans une large mesure, de la qualité et de l'actualité de ces données administratives.

Il est indispensable que le Bureau vérifie la qualité de toutes les données qu'il reçoit. À cette fin, nous avons mis au point et utilisé au cours des derniers recensements économiques des programmes de contrôle sur gros ordinateur. Toutefois, étant donné le très grand nombre de tableaux de vérification requis et compte tenu du fait que les ressources nécessaires à la

¹ James R. Jonas et Paul S. Hanczaryk, Economic Surveys Division, U.S. Bureau of the Census, Washington, D.C., U.S.A., 20233.

- EFFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Methods*. Philadelphia: SIAM.
- EFFRON, B. (1987). Better Bootstrap confidence intervals (avec discussion). *Journal of the American Statistical Association*, 82, 171-185.
- FELLEGI, I. P., et SUNTER, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.
- HALL, P. (1988). Theoretical comparison of Bootstrap confidence intervals. *Annals of Statistics*, 16, 927-953.
- KELLEY, R. P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A. P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B., et KENNEDY, J.M. (1962). Record linkage. *Communications of the Association for Computing Machinery*, 5, 563-566.
- NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., et ABBATT, J.D. (1983). Reliability of computerized versus manual searches in a study of the health of Eldorado Uranium workers. *Computers in Biology and Medicine*, 13, 157-169.
- SCHUREN, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-381.
- SCHUREN, F. (1985). Methodological issues in linkage of multiple data bases, dans *Record Linkage Techniques - 1985*, colligé par W. Alvey et B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 155-167.
- STATISTIQUE CANADA (1982). Record Linkage Software, Division des systèmes informatiques.
- STATISTIQUE CANADA (1983). Système itératif général de chaînage d'articles, Division des systèmes informatiques.
- STATISTIQUE CANADA (1984). Record Linkage Software, Division de la planification et du soutien en informatique.
- TEPPING, B. J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- U.S. DEPARTMENT OF AGRICULTURE (1979). List Frame Development: Procedures and Software, Statistical Reporting Service.
- U.S. BUREAU OF THE CENSUS (1978a). UNIMATCH: A Record Linkage System, Survey Research Division.
- U.S. BUREAU OF THE CENSUS (1978b). ZIPSTAN: Generalized Address Standardizer, Survey Research Division.
- WINKLER, W. E. (1985a). Preprocessing of lists and string comparison, dans *Record Linkage Techniques - 1985*, colligé par W. Alvey et B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 181-187.
- WINKLER, W. E. (1985b). Exact matching lists of businesses: Blocking, subfield identification, and Information Theory, dans *Record Linkage Techniques - 1985*, colligé par W. Alvey et B. Kilss, U. S. Internal Revenue Service, Publication 1299 (2-86), 227-241.
- WINKLER, W. E. (1985c). Exact matching lists of businesses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443.
- WINKLER, W. E. (1987). An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses, Energy Information Administration, rapport technique.
- WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

pour de produire des règles d'appariement meilleures que les règles qui n'utilisent pas ces mêmes comparaisons. Ce qui compte surtout lorsqu'on utilise des comparaisons additionnelles, c'est de savoir exploiter convenablement le pouvoir de différenciation additionnel qui en résulte. La série de comparaisons que nous avons faites dans cet article – notamment la comparaison des sous-zones de la zone réservée au nom – n'est pas indépendante au sens de l'équation (2.3). Le but premier de cette série de comparaisons est d'illustrer des méthodes qui permettent d'obtenir systématiquement de meilleures règles d'appariement lorsque l'hypothèse de l'indépendance conditionnelle n'est pas valide.

4.4 Autres critères de groupage

Lorsqu'on utilise une série de critères de groupage pour réduire le nombre de paires de $A \times B$ qui doivent faire l'objet d'un traitement plus poussé, on vise deux objectifs contradictoires. D'une part, on cherche à réduire fortement le nombre de paires qui doivent être traitées et à obtenir un ensemble (de paires) où les règles d'appariement peuvent départager clairement les concordances et les non-concordances. D'autre part, on cherche à obtenir un ensemble qui renferme autant d'éléments de M (concordances) que possible. Pour savoir s'il est souhaitable de définir des critères de groupage additionnels, il faut d'abord obtenir des estimations du nombre de concordances qui échappent à une série donnée de critères de groupage. Si les estimations sont raisonnablement faibles, il ne sera pas nécessaire de définir de nouveaux critères.

Pour estimer le nombre de concordances qui échappent à des séries données de critères, Scheuren (1983) propose d'utiliser les méthodes régulières de saisie-ressaisie telles qu'elles sont décrites dans Bishop, Fienberg et Holland (1975, chapitre 6). Winkler (1987) a appliqué ces méthodes aux données empiriques et aux quatre séries de critères de groupage présentées dans cet article. Le modèle logarithmique du meilleur ajustement pour le nombre d'enregistrements saisis et d'enregistrements non-saisis selon les quatre séries de critères de groupage a servi à calculer un intervalle de confiance pour le nombre de concordances omises. Suivant l'hypothèse de la normalité asymptotique, on a calculé un intervalle de 95 % (27,160). Cet intervalle représente entre 1 et 5 % de toutes les concordances.

5. RÉSUMÉ

Les résultats exposés dans cet article montrent que l'hypothèse de l'indépendance conditionnelle n'est pas toujours valide. Lorsque l'hypothèse n'est pas vérifiée, on peut élaborer des règles d'appariement qui sont supérieures à la règle normale. Selon un intervalle fixe pour les taux d'erreur, les nouvelles règles réduisent la taille de la région des cas indéterminés.

BIBLIOGRAPHIE

- BISHOP, Y. M. M., FIENBERG, S. E., et HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.
- COCHRAN, W. G., et COX, G. M. (1957). *Experimental Designs*. New York: John Wiley and Sons.
- EFFRON, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics* 7, 1-26.

L'élément σ^1 peut désigner le prénom, l'élément σ^2 le numéro civique, l'élément σ^3 l'âge et ainsi de suite.

Pour chaque σ , nous pouvons désigner $P(\sigma^i = \sigma^0 | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, M)$ comme la **composante de discrimination successive additionnelle conditionnelle** de σ^i dans M , $i = 1, 2, \dots, K$. Ces composantes dépendent de $\sigma^1, \sigma^2, \dots, \sigma^K$ qui sont le résultat d'un reclassement. Les éléments du membre de droite de l'équation (4.1) sont indépendants les uns des autres. Nous pouvons aussi définir des composantes semblables dans U .

Un reclassement vise essentiellement à considérer un ensemble déterminé de probabilités conditionnelles pour $\gamma \in \Gamma$. En ce qui a trait au reclassement simple, nous faisons varier $\sigma = \sigma(\gamma)$ dans $\sigma(\Gamma)$ puisque $\gamma \in \Gamma$. Ainsi, pour tous $\sigma \in \sigma(\Gamma)$,

$$W \equiv W(\gamma) = \text{Log}_2[m(\gamma)/n(\gamma)]$$

$$= A^1 + A^2 + \dots + A^K, \tag{4.2}$$

$$\text{où } A^i \equiv \text{Log}_2[P(\sigma^i = \sigma^0 | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, M)/P(\sigma^i = \sigma^0 | \sigma^1, \sigma^2, \dots, \sigma^{i-1}, U)] \text{ pour } i = 1, 2, \dots, K.$$

La seconde série de faits concerne des transformations qui appliquent le rapport R défini en (2.1) sur des nombres réels que nous appelons **poids**. Pour chaque paire d'erreurs de type I et de type II, nous considérons une transformation qui place les poids se rapportant à des liens dans l'intervalle le plus élevé, les poids se rapportant à des non-liens dans l'intervalle le plus bas et les poids se rapportant à des cas indéterminés dans l'intervalle situé entre le plus bas et le plus élevé. Cette transformation produit des règles qui peuvent s'énoncer comme celles définies en (2.2) et qui équivalent à la règle de Fellegi-Sunter, étant donné le même intervalle fixe pour le taux d'erreur. S'il s'agit d'une transformation monotone, les nouveaux poids donnent des règles équivalentes à la règle de Fellegi-Sunter originale pour tous les niveaux d'erreur.

La correction de poids par la méthode de la plus grande ascension détermine implicitement une transformation du rapport R ainsi qu'un reclassement simple qui est fixe pour tous les $\gamma \in \Gamma$ et le même dans M et U . Le fait que les poids soient corrigés indépendamment les uns des autres garantit un reclassement unique. Les poids corrigés $W' \pm c_i$ sont des estimations qui remplacent W' dans l'expression (2.3) pour une constante réelle c_i , $i = 1, 2, \dots, k$.

Le fait que l'on obtient des régions de cas indéterminés plus petites avec les poids corrigés signifie que les nouveaux poids totaux reflètent plus fidèlement une transformation du Log_2 du rapport des probabilités réelles définies par le membre de gauche de l'équation (4.1), étant donné un intervalle fixe pour les taux d'erreur. Les nouveaux poids totaux représentent des estimations qui transforment le membre de droite de l'équation (4.2).

La méthode de correction nous permet de mieux utiliser le pouvoir de différenciation additionnel d'une zone étant donné une autre zone, d'une troisième zone étant donné les deux premières, et ainsi de suite. Notons qu'il n'est pas nécessaire de connaître la transformation ou l'effet conditionnel précis qui découle du reclassement.

La méthode de correction ressemble aux nouvelles méthodes «bootstrap» (Efron 1987; Hall 1988). Ces nouvelles méthodes seront utiles dans la mesure où il existera des transformations monotones, des constantes de biais et des constantes d'accélération qui feront correspondre parfaitement les intervalles de confiance des distributions originales aux intervalles de confiance de distributions normales particulières. Il ne sera pas nécessaire de connaître ces transformations et ces constantes.

4.3 Utilité des comparaisons qui dépendent d'autres comparaisons

Intuitivement, il est justifié de faire un certain nombre de comparaisons, dont certaines peuvent dépendre en partie d'autres comparaisons, parce qu'elles sont susceptibles de créer un

4. ANALYSE

Cette section se divise en quatre parties. Dans la première, nous analysons la robustesse des corrections faites par la méthode de la plus grande ascension. Ensuite, nous décrivons le genre de conditions implicites qu'imposent ces corrections. Dans la troisième partie, nous considérons l'utilité de faire des comparaisons qui dépendent en partie d'autres comparaisons. Enfin, dans la quatrième partie, nous décrivons des méthodes qui permettent de définir de nouvelles séries de critères de groupage.

4.1 Robustesse des corrections faites par la méthode de la plus grande ascension.

La taille des régions de cas indéterminés est assez sensible à la série de poids qui sont modifiées par la méthode de la plus grande ascension. Il y a deux cas d'application (dont l'un a été exposé dans cet article) où le nombre de cas indéterminés est environ 100 et deux autres où il est environ 200. Si l'on considère globalement les quatre cas d'application, on constate que la méthode de la plus grande ascension produit un moins grand nombre de cas indéterminés que la meilleure méthode qui puisse lui être substituée (où on trouve 700 cas indéterminés). Les poids modifiés par la méthode de la plus grande ascension l'ont été à des degrés très divers d'un cas à l'autre. Jamais plus que 8 poids sur 30 ont été modifiés.

Il est raisonnable de supposer que la méthode de pondération par la plus grande ascension donnera de meilleurs résultats lorsque l'hypothèse de l'indépendance conditionnelle sera difficilement acceptable. Dans trois des quatre cas d'application, on n'a pas utilisé de signification fondée sur la «bootstrap» pour vérifier l'hypothèse.

Il devrait être facile d'obtenir de petits échantillons qui permettent de faire des corrections comme celles qui ont été faites dans cet article. Un échantillon de taille 100 devrait suffire pour chaque classe. L'échantillon utilisé dans la section 3.3 pour l'application de la méthode «bootstrap» avait une taille approximative de 100 pour chaque classe. Des échantillons de taille $n = 30$ ou $n = 50$ pour chaque classe étaient trop petits pour que les résultats de la méthode «bootstrap» n'indiquent des progrès quantifiables. Des échantillons de taille $n = 200$ produisaient à peu près les mêmes intervalles de confiance «bootstrap» que des échantillons de taille $n = 100$. De nombreux systèmes d'appariement d'enregistrements (p. ex.: Département de l'agriculture des E.-U., 1979, Département du commerce des E.-U., 1978a, Statistique Canada 1984) permettent de modifier les paramètres d'appariement à partir de données d'échantillon. La réestimation des paramètres à l'aide de données d'échantillon est une fonction puissante du système itératif général de chaînage d'articles de Statistique Canada (1983). En règle générale, ce genre de fonction consiste à réestimer directement les probabilités marginales $m_i(\gamma')$ et $u_i(\gamma')$. Elle ne consiste pas à modifier des poids comme cela a été fait dans cet article.

4.2 Genre d'effet conditionnel représenté par les poids modifiés

Pour les besoins de cette analyse, nous devons poser deux séries de faits. La première série concerne le pouvoir de différenciation conditionnel des éléments de γ . Soit σ un vecteur ayant pour éléments $\sigma_1, \sigma_2, \dots, \sigma_K$ ce vecteur est le résultat d'un reclassement des éléments $\gamma_1, \gamma_2, \dots, \gamma_K$ de γ .

Alors

$$P(\gamma | M) = P(\sigma | M) =$$

$$P(\sigma_1 = \sigma_1^0, \sigma_2 = \sigma_2^0, \dots, \sigma_K = \sigma_K^0 | M) =$$

$$P(\sigma_1 = \sigma_1^0 | M) \cdot P(\sigma_2 = \sigma_2^0 | \sigma_1, M) \dots P(\sigma_K = \sigma_K^0 | \sigma_1, \sigma_2, \dots, \sigma_{K-1}, M).$$

(4.1)

Voici un exemple d'une fausse non-concordance déduite à l'aide de poids de type C.

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Johns Geo M	167 Sycamore St	Springfield	OH	53315
Geo M Johns Jobber	167 Sycamore	Spring Field	OH	53315.

À cause de l'insertion ou de la suppression de blancs dans des zones correspondantes, l'ordre-nature a tendance à définir les paires d'enregistrements en question comme des non-concordances.

3.3 Application de la méthode «bootstrap»

Les résultats présentés dans cette sous-section découlent de l'application de méthodes de calcul d'intervalles de confiance «bootstrap» de complexité croissante (tableau 9). Pour chaque classe, on utilise 500 échantillons répétés pour calculer des intervalles de confiance de 90% pour les estimations du nombre d'enregistrements définis comme des cas indéterminés. Le taux d'erreur de classification maximum est fixé à 5%.

Le premier genre d'intervalle est l'intervalle «bootstrap» ordinaire qui repose en partie sur la théorie normale (Efron 1979). Le second genre d'intervalle, désigné par BC, est un intervalle qui comporte un ajustement pour le biais (Efron 1979, 1982). Le troisième genre d'intervalle, désigné par BC_a, est déterminé au moyen d'un ajustement à constante d'accélération pour le biais et l'asymétrie (Efron 1987; aussi Hall 1988).

En examinant le tableau 9, nous constatons que les trois intervalles sont à peu près de la même longueur pour une classe donnée. Si la méthode de correction utilisée pour obtenir des poids de type C dépendait largement des échantillons de référence, nous nous attendrions que les intervalles de confiance rattachés aux poids de type C soient plus grands que ceux rattachés aux poids de type U.

Le fait que l'on observe de grands intervalles d'un côté comme de l'autre indique que les résultats dépendent largement des échantillons de référence. En outre, le fait que les intervalles de confiance ordinaires soient comparables aux intervalles BC et BC_a correspondants indique que les distributions respectives ne sont ni biaisées ni asymétriques. Le nombre de cas indéterminés contenus dans les intervalles fondés sur les poids de type C est presque toujours inférieur au nombre de cas indéterminés contenus dans les intervalles fondés sur les poids de type U. Seuls les intervalles calculés pour les classes 3 et 4 montrent un léger chevauchement. Il est donc raisonnable d'accepter l'hypothèse selon laquelle la règle d'appariement fondée sur des poids de type C surpasse constamment la règle fondée sur des poids de type U.

Tableau 9

Intervalles de confiance «bootstrap» de 90% pour le nombre de cas indéterminés
500 échantillons répétés

Type de poids	Classe	Intervalle ordinaire	Intervalle BC	Intervalle BC _a
C	1	(42,117)	(37,108)	(37,108)
C	2	(0, 0)	(7, 7)	(7, 7)
C	3	(31,154)	(34,156)	(34,156)
C	4	(0, 36)	(0, 39)	(0, 39)
U	1	(122,192)	(128,196)	(128,196)
U	2	(383,501)	(383,501)	(383,501)
U	3	(149,201)	(142,197)	(142,197)
U	4	(35, 82)	(33, 81)	(33, 81)

Tableau 8
Correction (par la plus grande ascension) des poids de concordance pour les sous-zones tirées de la zone LM-NOM¹

Sous-zones				
Classe	1	2	3	4
1	.	.	—	+
2	++	++	+	+
3	+	+	—	++
4	.	+	—	+

¹ « . » signifie un écart inférieur à 1.0, « + » et « — » signifient un écart supérieur à 1.0 et inférieur à 2.5 « ++ » signifie un écart supérieur à 2.5.

La règle d'appariement fondée sur des poids de type C se distingue de celle fondée sur des poids de type U à deux points de vue. D'une part, nous modifions les poids de concordance qui se rattachent aux quatre sous-zones du NOM après avoir classé les mots par ordre décroissant de longueur (tableau 8). Les seules variations appréciables (supérieures à 2.5 sur l'échelle log₂) sont observées dans la classe 2.

D'autre part, nous n'utilisons le poids de concordance que si quatre sous-zones correspondent, soit les trois sous-zones de VILLE et la sous-zone ETAT, concordent. De fait, la variation de poids accroît généralement le pouvoir de différenciation **relatif** des concordances/non-concordances dans toutes les sous-zones sauf celles de la zone VILLE.

La plus forte réduction du nombre de cas indéterminés (de 409 à 0) est observée dans la classe 2. Les non-liens qui ont une zone VILLE conforme sont proportionnellement légèrement plus nombreux que les liens (.95 ≈ 359/379) contre (.91 ≈ 223/245).

Nous donnons ci-dessous l'exemple d'une concordance qui n'est pas reconnue comme un lien selon la règle fondée sur des poids de type U mais qui l'est selon la règle fondée sur des poids de type C.

NOM	RUE	VILLE	ETAT	CODE POSTAL
Robertis Heat Oils	167 Sycamore St	Dayton	OH	53315
Maxwell S Robert Heat Oil	167 Sycamore St	Dayton	OH	53315.

Les six premiers chiffres du numéro de téléphone concordent également. Voici un exemple d'une fausse concordance déduite à l'aide de poids de type C.

NOM	RUE	VILLE	ETAT	CODE POSTAL
Molar Petro	167 Sycamore St	Dayton	OH	53315
Petrochem	167 Sycamore St	Dayton	OH	53315.

Ces deux entreprises sont situées à la même adresse et ont le même numéro de téléphone.

Tableau 5

Taux d'erreur et nombre de cas indéterminés pour diverses méthodes de pondération				
Type de poids	Proportion des concordances		Total des paires	
	classées par		classées	
	Erreur dans les non-concordances	Erreur dans les concordances	Non-concordances	Concordances
AA	.047	.020	964	2009
A	.041	.015	952	2481
U	.050	.020	1083	2707
C	.033	.019	1441	2947
				97
				1512
				1052
				695

Tableau 6

Résultats de l'application d'une règle d'appariement fondée sur des poids de type U pour définir les concordances et les non-concordances (Taux d'erreur de classification global de 5%)						
Classe	Poids limites		Classes par erreur dans les		Total des paires classées	
	LOWER	UPPER	Non-concor-	Concor-	Non-concor-	Concor-
	dances	dances	dances	dances	dances	dances
1	0.5	6.5	39	14	674	264
2	-4.5	3.5	2	4	100	115
3	-4.5	6.5	2	1	55	42
4	2.5	11.5	11	2	254	46
Total			54	21	1083	467
					695	
					2245	

Tableau 7

Résultats de l'application d'une règle d'appariement fondée sur des poids de type C pour définir les concordances et les non-concordances (Taux d'erreur de classification global de 3%)						
Classe	Poids limites		Classes par erreur dans les		Total des paires classées	
	LOWER	UPPER	Non-concor-	Concor-	Non-concor-	Concor-
	dances	dances	dances	dances	dances	dances
1	4.5	7.5	28	8	692	274
2	2.5	2.5	5	3	379	245
3	-0.5	4.5	9	4	266	78
4	8.5	8.5	47	21	1441	707
Total						
					97	
					2245	

Les calculs et les corrections doivent être effectués uniformément pour tous les échantillons de référence. Les poids individuels corrigés, les poids totaux et les bornes doivent tous être obtenus à l'aide des mêmes méthodes de correction. Si un poids individuel est corrigé à la hausse (étape 2) d'une quantité x ou d'un pourcentage y pour un échantillon, la même correction doit s'appliquer pour les autres échantillons.

Comme les distributions sous-jacentes peuvent n'être pas normales ou être biaisées et asymétriques, nous pouvons utiliser de nouvelles méthodes proposées par Efron (1982, 1987; aussi Hall, 1988) pour déterminer les intervalles de confiance. Hall (1988) a démontré la valeur théorique de la méthode «bootstrap» non paramétrique, qui renferme un mécanisme d'ajustement à constante d'accélération pour les distributions asymétriques.

3. RÉSULTATS

Cette section se divise en trois parties. La première est une comparaison générale des quatre méthodes de pondération décrites dans la sous-section 2.3.5. La seconde approfondit les deux meilleures méthodes parmi ces quatre. Enfin, la troisième contient les résultats de l'évaluation par la méthode «bootstrap».

3.1 Comparaison générale

Nous fixons une borne supérieure de 5% pour la proportion de concordances classées par erreur dans les non-concordances et une borne supérieure de 2% pour la proportion de non-concordances classées par erreur dans les concordances. Comme nous utilisons des données discrètes, les taux d'erreur réels n'égaleront généralement pas les bornes supérieures (tableau 5, colonnes 2 et 3).

Nous constatons qu'à mesure que s'accroît la complexité de la méthode de pondération, le nombre de cas indéterminés (taille de la région soumise à une révision manuelle) diminue de façon appréciable, passant de 1,512 à 97. Cela indique que la complexité accrue des méthodes de calcul des poids engendre de meilleures règles de décision.

Nous constatons aussi que les deux dernières méthodes, qui consistent à calculer séparément des poids de comparaison individuels dans les quatre premières classes, produisent les plus petits ensembles de cas indéterminés (695 et 97 respectivement).

3.2 Les meilleures méthodes

Considérons plus en détail les deux meilleures méthodes, soit les règles d'appariement fondées sur les poids de type U et de type C. Les résultats obtenus par l'application des poids de type U et de type C sont présentés dans les tableaux 6 et 7 respectivement. En déterminant les poids limites pour chaque classe, nous fixons une limite supérieure approximative de 5% pour la proportion de non-concordances classées par erreur dans les concordances et une limite supérieure approximative de 2% pour la proportion de concordances classées par erreur dans les non-concordances. La borne supérieure générale est conservée.

Si nous comparons les colonnes 4 et 5 des tableaux 6 et 7, nous constatons que les totaux correspondants sont comparables. Ce résultat est conforme à la méthode de délimitation. Dans chaque classe, la règle d'appariement fondée sur des poids de type C produit un moins grand nombre de cas indéterminés que la règle d'appariement fondée sur des poids de type U.

De fait, le nombre d'enregistrements considérés comme cas indéterminés est moindre pour les classes 1 et 4 (83 contre 55 et 44 contre 0 respectivement) et considérablement moindre pour les classes 2 et 3 (409 contre 0 et 159 contre 42 respectivement).

La méthode fondée sur des poids de type C permet de classer la totalité des paires continues dans les classes 2 et 4.

Tableau 4
Règles d'appariement (selon le mode de calcul des poids) et ensembles de paires auxquels elles s'appliquent

Type	Calcul de poids individuels	Règle d'appariement
AA	Appliqué uniformément à toutes les paires contenues dans les 5 classes	S'applique à toutes les paires
A	Appliqué uniformément à toutes les paires contenues dans les classes 1 à 4	Identifie les paires de la classe 5 comme des liens; applique la règle de Fellegi-Sunter aux paires des 4 autres classes
U	Appliqué uniformément dans chacune des 4 premières classes	Identifie les paires de la classe 5 comme des liens; applique la règle de Fellegi-Sunter à chacune des paires des 4 premières classes
C	Appliqué uniformément dans chacune des 4 premières classes	Même que U, mais modifie les poids à l'aide de la méthode de la plus grande ascension.

2.4 Méthodes d'évaluation

La méthode d'évaluation fondamentale consiste à suivre l'évolution de la taille de la région des cas indéterminés lorsqu'on applique les divers genres de règles d'appariement suivant un intervalle de taux d'erreur fixe.

La méthode «bootstrap» d'Efron (1987, 1982, 1979) permet d'estimer des intervalles de confiance pour des paramètres statistiques comme le nombre de cas indéterminés. Comme ces paramètres statistiques sont déterminés suivant des règles complexes, il est peu probable que l'on puisse établir des estimations de ces paramètres en forme analytique.

S'il y a des ensembles de paires pour lesquelles l'authenticité ou la fausseté des concordances est vérifiée, nous pouvons utiliser la méthode «bootstrap» d'Efron pour estimer la variation des paramètres selon les règles suivantes:

1. Tirer (avec remise) des échantillons de référence de même taille.
 2. Estimer les poids de comparaison individuels définis en (2.4) en se servant des renseignements relatifs à l'authenticité ou à la fausseté des concordances dans l'échantillon et utiliser ces poids pour estimer le poids total par la formule (2.3).
 3. Calculer les bornes LOWER et UPPER à l'aide de chaque échantillon (en l'occurrence, nous limitons la proportion de liens classés comme non-concordances à 2% et la proportion de non-liens classés comme concordances à 3%).
 4. À l'aide des poids de comparaison individuels estimés à l'étape 2, calculer un poids de comparaison total pour chaque paire contenue dans l'ensemble prélevé. Utiliser les bornes calculées à l'étape 3 pour identifier les paires comme liens, cas indéterminés ou non-liens.
 5. À l'aide des estimations d'échantillons, déterminer la moyenne et la variance des poids limites, des taux d'erreur de classification et du nombre de cas indéterminés.
- Les bornes (2 et 3%, étape 3) visent à faire en sorte que les taux d'erreur de classification pour toute la base de données soient inférieurs à 5%.

Tableau 2

Critères de groupage	
#	Caractères utilisés
1.	CODE POSTAL 3 chiffres, NOM 4 caractères
2.	CODE POSTAL 5 chiffres, RUE 6 caractères
3.	TÉLÉPHONE 10 chiffres
4.*	Zone LM-NOM; ensuite, appliquer le critère 1.
* Ce critère renferme un mécanisme de suppression qui empêche l'appariement avec des mots qui reviennent souvent comme «OIL», «FUEL», «CORP», et «DISTRIBUTOR.»	

Tableau 3

Ensembles de paires formées à l'aide des critères de groupage		
Classe	n° de paires	Critères de groupage déterminants
1	1021	conformes au critère n° 1 et à aucun autre ou conformes aux critères 1 et 4 et à aucun autre.
2	624	conformes au critères n° 2 et à aucun autre ou conformes aux critères 2 et 3 et à aucun autre.
3	256	conformes au critère n° 3 seulement.
4	344	conformes au critère n° 4 seulement.
5	2240	conformes à au moins un critère mais n'appartenant à aucune des classes précédentes.

Les seules paires considérées sont celles qui répondent à au moins un des critères de groupage du tableau 2.

Nous divisons en cinq classes (tableau 3 ci-dessous) l'ensemble de paires formées à l'aide des quatre séries de critères de groupage.

La classe 5 renferme des paires qui répondent habituellement à au moins deux critères de groupage. Les cinq classes englobent 2,991 concordances et 1,494 non-concordances et auraient dû comprendre 59 autres concordances reconnues comme telles. Winkler (1985b, 1987) examine en détail la formation des classes et la définition des séries de critères de groupage.

Les règles d'appariement sont classées selon les méthodes de calcul des poids de comparai-son individuels et la manière dont les nouvelles règles d'appariement sont définies.

Le premier type de calcul de poids (AA) est un agrégat qui porte sur toutes les paires. Le second type de calcul (A) est une agrégat qui porte sur les 4 premières classes. Le troisième (U) produit des valeurs distinctes à l'intérieur des 4 premières classes. Le quatrième (C) utilise la méthode de la plus grande ascension pour modifier le calcul du type U.

À mesure que sont définies successivement les règles d'appariement, le calcul des poids devient de plus en plus complexe. Les concordances qui ne sont pas incluses dans l'une ou l'autre des 5 classes ne sont pas considérées dans la section des résultats parce que leur nombre est le même pour chacune des quatre règles d'appariement.

poids totaux obtenus par cette correction donnaient des ensembles de liens possibles beaucoup plus petits suivant un intervalle fixe pour les taux d'erreur, cela pourrait mettre en doute la robustesse du modèle de Fellegi-Sunter.

2.3.4 Méthodes générales de correction

Deux méthodes générales de correction se rattachent aux méthodes de calcul des poids de comparaison individuels. La première consiste à subdiviser en plusieurs parties le sous-ensemble de paires de $A \times B$ pour lequel des poids de comparaison individuels sont calculés. On détermine la règle d'appariement en faisant en sorte que la règle fondamentale de Fellegi-Sunter corresponde uniquement aux divers sous-ensembles pour lesquels des poids sont calculés. Les poids de comparaison individuels peuvent varier largement selon les sous-ensembles.

La seconde méthode consiste à modifier les poids de comparaison individuels. Dans l'hypothèse de l'indépendance, considérons l'équation:

$$W = \text{Log}_2(P(\gamma \in B_1 \cap B_2 \cap \dots \cap B_K | M) / P(\gamma \in B_1 \cap B_2 \cap \dots \cap B_K | U))$$
$$= W^1 + W^2 + \dots + W^K,$$

où, pour $i = 1, 2, \dots, K$, $W^i = \text{Log}_2(P(\gamma \in B_i | M) / P(\gamma \in B_i | U))$ et B_i est l'ensemble $\{\gamma^i = \gamma_0^i\}$ ou son complément. Nous voulons trouver des méthodes flexibles qui permettent de corriger les W^i , $i = 1, 2, \dots, K$, de telle sorte que leur somme donne de meilleures règles d'appariement.

S'il existe un échantillon pour lequel l'authenticité ou la fausseté des concordances a été vérifiée, nous pouvons alors estimer les poids de comparaison individuels (Tepping 1968) et les corrections.

La méthode de correction la plus simple est la méthode de la plus grande ascension (voir, par exemple, Cochran et Cox 1957). Nous commençons par utiliser les renseignements relatifs à l'authenticité ou à la fausseté des concordances dans un échantillon afin d'estimer des probabilités comme celles définies en (2.4). Ces probabilités servent ensuite à calculer des poids de comparaison individuels, que l'on additionne afin d'obtenir une estimation du poids total (2.3). On peut déterminer les bornes UPPER et LOWER de l'expression (2.2) pour chaque intervalle de taux d'erreur de type I ou de type II qui est fixe. On en déduit aussitôt le nombre de liens possibles pour des règles comme celle définie en (2.2).

Dans un deuxième temps, nous choisissons un poids de comparaison individuel, que nous modifions d'une valeur fixe (par exemple ± 1), nous recalculons le poids total de (2.3) au moyen du nouveau poids individuel, puis nous déterminons les nouvelles bornes UPPER et LOWER ainsi qu'une nouvelle région de liens possibles.

Si, avec un intervalle fixe, la taille de la région des liens possibles diminue, nous corrigeons (à la hausse ou à la baisse) le poids de comparaison individuel jusqu'à ce que la taille cesse de décroître. Nous poursuivons en modifiant les autres poids individuels de la même façon.

Si la taille de la région des cas indéterminés décroît de façon appréciable, alors nous savons que l'hypothèse de l'indépendance conditionnelle n'est pas valide pour l'ensemble des comparaisons. Si cette hypothèse était valide, les poids estimés reflèteraient fidèlement les poids réels. La taille de la région des cas indéterminés serait minimum en vertu du théorème de Fellegi-Sunter. Une règle d'appariement qui repose sur des poids de comparaison individuels corrigés dépend de l'échantillon utilisé pour la méthode de la plus grande ascension.

2.3.5 Méthodes particulières

Nous avons besoin de renseignements supplémentaires pour décrire les méthodes particulières par lesquelles nous calculons les poids et définissons les règles d'appariement correspondantes appliquées dans cet article.

Cette hypothèse revient essentiellement à dire que le fait qu'il y ait concordance du nom de famille est indépendant du fait qu'il y ait concordance d'autres éléments comme le numéro civique ou l'âge.

La seconde façon de simplifier le calcul est d'utiliser une fonction du rapport défini en (2.1), qui se prête bien à des calculs. Nous avons choisi la fonction Log_2 . Nous avons donc

$$W \equiv W(\gamma) = \text{Log}_2[m(\gamma)/n(\gamma)]$$

$$(2.3) \quad = W^1 + W^2 + \dots + W^K,$$

où $W^i \equiv \text{Log}_2[m_i(\gamma^i)/n_i(\gamma^i)]$ pour $i = 1, 2, \dots, K$. Nous appelons W le poids de comparaison total lié à une paire et W^i , $i = 1, 2, \dots, K$, les poids de comparaison individuels. Pour le reste de cet article, nous supposons que chaque composante γ^i , $i = 1, 2, \dots, K$, de γ est une variable binaire (par exemple concordance/non-concordance). Pour des raisons de commodité, nous désignons les cas de concordance pour la i -ème composante par γ_o^i , $i = 1, 2, \dots, K$. Suivant l'hypothèse de l'indépendance conditionnelle, pour chaque $i = 1, \dots, K$, nous devons estimer des probabilités du genre

$$P(\gamma = \gamma_o^i | M) \text{ et } P(\gamma = \gamma_o^i | U).$$

(2.4)

Si nous avons un ensemble de paires pour lesquelles l'authenticité ou la fausseté des concordances est vérifiée, nous divisons cet ensemble, pour chaque cas de concordance γ_o^i , $i = 1, 2, \dots, K$, de manière à obtenir les quatre sous-ensembles définis par (2.4) avant de procéder à l'estimation.

En l'absence de l'hypothèse d'indépendance conditionnelle, nous devons estimer $2: (2^K - 1)$ probabilités définies suivant la formule (2.1) et diviser l'ensemble de paires pour lesquelles l'authenticité ou la fausseté des concordances est vérifiée en $2: (2^K - 1)$ sous-ensembles. Même avec un petit nombre de comparaisons (disons six au maximum), nous pourrions ne pas être en mesure d'obtenir des échantillons suffisamment grands pour estimer avec précision les probabilités.

2.3.3 Validité de l'hypothèse de l'indépendance conditionnelle

Winkler (1985c) a montré que l'hypothèse de l'indépendance conditionnelle n'est pas valide pour des comparaisons simples de portions des zones NOM et RUE pour des listes d'entreprises. En utilisant les mêmes portions des zones NOM et RUE, Kelley (1986) a montré que l'hypothèse de l'indépendance conditionnelle n'était pas valide pour des fichiers de personnes. De plus, Kelley et Winkler ont montré, chacun de son côté, que l'efficacité d'appariement dépendait beaucoup de l'ensemble de paires pour lequel des probabilités comme celles définies en (2.4) étaient calculées.

Fellegi et Sunter précisent que si l'hypothèse de l'indépendance conditionnelle n'est pas valide, on ne pourra plus interpréter d'une manière strictement probabiliste les estimations des poids calculés selon la formule (2.3). Autrement dit, la règle d'appariement du théorème de Fellegi-Sunter pourrait ne pas réduire au minimum le nombre de cas indéterminés. Néanmoins, Fellegi et Sunter croient à la robustesse de leur modèle si l'on s'écarte de l'hypothèse de l'indépendance.

Suivant l'hypothèse de l'indépendance, la probabilité équivalente au produit de probabilités comme celles définies en (2.4). Si nous avons un ensemble de paires pour lesquelles l'authenticité ou la fausseté des concordances est vérifiée, il est alors possible de corriger les probabilités de l'expression (2.4) pour les cas où l'hypothèse de l'indépendance ne tiendrait plus. Si les

2.3. Méthodes de calcul

Cette section se divise en cinq parties. La première contient une description de la règle d'appariement générale du modèle de Fellegi-Sunter. La seconde contient une description de la version simplifiée des méthodes de calcul lorsqu'une hypothèse d'indépendance conditionnelle est posée. Dans la troisième partie, nous discutons de la validité de l'hypothèse d'indépendance conditionnelle. La quatrième partie expose deux méthodes générales permettant de modifier les méthodes de calcul. Enfin, la cinquième partie contient une description des méthodes de calcul utilisées précisément pour cet article.

2.3.1 Définition générale de la règle d'appariement

Pour comprendre pourquoi on utilise des méthodes de calcul particulières, considérons le rapport de vraisemblance suivant

$$(2.1) \quad R \equiv R[\gamma(a,b)] = m(\gamma)/u(\gamma).$$

Nous remarquons que si γ représente une comparaison de K zones, nous avons au moins 2^K probabilités de forme $m(\gamma)$. Si γ indique une concordance pour K zones, il serait normal d'observer cela plus souvent pour l'ensemble des concordances M que pour l'ensemble des non-concordances U . On aurait alors un rapport R élevé. Par ailleurs, si γ indique des cas de non-concordance, le rapport R sera peu élevé. Si, dans l'équation ci-dessus, le numérateur est positif et le dénominateur nul, nous attirerons l'attention sur une valeur arbitraire très élevée. La règle d'appariement de Fellegi-Sunter s'énonce alors comme suit:

- Si $R > \text{UPPER}$, (a,b) est définie comme un lien.
 Si $\text{LOWER} \leq R \leq \text{UPPER}$, (a,b) est définie comme un cas indéterminé (2.2)
 Si $R < \text{LOWER}$, (a,b) est définie comme un non-lien.

Les bornes LOWER et UPPER sont déterminées par l'intervalle de taux d'erreur voulu.

2.3.2 Simplification suivant l'hypothèse de l'indépendance conditionnelle

En pratique, on simplifie le calcul de deux façons. La première consiste à poser l'hypothèse de l'indépendance conditionnelle de Fellegi et Sunter (1969):
 Pour chaque $\gamma \in \Gamma$

$$m(\gamma) = m_1(\gamma_1) \cdot m_2(\gamma_2) \cdot \dots \cdot m_K(\gamma_K) \text{ et}$$

$$u(\gamma) = u_1(\gamma_1) \cdot u_2(\gamma_2) \cdot \dots \cdot u_K(\gamma_K)$$

où, pour $i = 1, 2, \dots, K$

$$m_i(\gamma_i) = P(\gamma_i | (a,b) \in M) \text{ et}$$

$$u_i(\gamma_i) = P(\gamma_i | (a,b) \in U).$$

Nous observons un vecteur d'information $\gamma(a,b)$ associé à la paire (a,b) et voulons identifier une paire comme lien (décision A_1), cas indéterminé (décision A_2) ou non-lien (décision A_3). Nous définissons une **règle d'appariement** L comme une application de Γ , l'espace des comparaisons, sur un ensemble de fonctions de décision aléatoires $D = \{d(\gamma) \text{ où}$

$$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}; \gamma \in \Gamma$$

et

$$\sum_3^{i=1} P(A_i|\gamma) = 1.$$

Deux types d'erreur sont possibles avec une règle d'appariement. Il se produit une **erreur de type I** lorsqu'une paire de l'ensemble des non-concordances est identifiée par erreur comme lien. La probabilité de cette erreur est définie

$$P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1|\gamma)$$

Il se produit une **erreur de type II** lorsqu'une paire de l'ensemble des concordances est identifiée par erreur comme non-lien. La probabilité de cette erreur est définie

$$P(A_3|U) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3|\gamma).$$

Fellegi et Sunter (1969) définissent une règle d'appariement L_0 , optimale avec les décisions correspondantes A_1, A_2 , et A_3 , l'optimalité de cette règle s'entend de la façon suivante:

THÉORÈME (Fellegi-Sunter 1969). Soit L' une règle d'appariement avec les décisions correspondantes A_1, A_2 , et A_3 de sorte que $P(A_3|M) = P(A_3|U)$ et $P(A_1|M) = P(A_1|U)$ (mêmes probabilités d'erreur qu'avec L_0). Alors, L_0 optimale du fait que $P(A_2|U) \leq P(A_2'|U)$ et $P(A_2|M) \leq P(A_2'|M)$.

Autrement dit, si L' et L_0 sont deux règles concurrentes qui ont les mêmes taux d'erreur de type I et de type II (lesquels sont des probabilités conditionnelles), alors la probabilité conditionnelle (pour l'ensemble U ou l'ensemble M) que l'on ne prenne pas de décision selon la règle L' est toujours supérieure à la probabilité conditionnelle que l'on ne prenne pas de décision selon la règle L_0 . L_0 est décrite dans la sous-section 2.3.1.

De fait, la règle d'appariement de Fellegi-Sunter est optimale par rapport à n'importe quel ensemble \tilde{Q} de paires ordonnées dans $A \times B$ si nous définissons des probabilités d'erreur $P_{\tilde{Q}}$ et une règle d'appariement $L_{\tilde{Q}}$ qui dépendent de \tilde{Q} . Ainsi, nous pourrions peut-être définir des sous-ensembles de $A \times B$ auxquels nous appliquerions divers genres de renseignements en quantités variables.

Par exemple, si nous avons un ensemble de paires pour lesquelles le numéro de téléphone est indiqué, nous pourrions utiliser le numéro de téléphone et quelques caractères du nom pour définir les liens. Pour d'autres paires, nous pourrions devoir utiliser aussi les renseignements contenus dans les zones RUE et VILLE.

Des critères de groupage sont souvent à l'origine des ensembles \tilde{Q} de paires ordonnées auxquelles est appliquée la règle d'appariement de Fellegi-Sunter. Les critères de groupage sont des indicatifs de tri qui servent à réduire le nombre de paires considérées. Au lieu de considérer toutes les paires contenues dans $A \times B$, nous pourrions considérer seulement les paires dont les éléments ont en commun les trois premiers chiffres du code postal ou l'abréviation du nom de famille (pour autant que cette abréviation soit acceptable).

Tableau 1

Sous-zones correspondantes comparées à raison d'un caractère à la fois	
ZONE	Colonnes de sous-zones
NOM	1-4, 5-10, 11-20, 21-30
RUE	1-6, 7-15, 16-30
CODE POSTAL	1-3, 4-5
VILLE	1-5, 6-10, 11-15
ÉTAT	1-2
TÉLÉPHONE	1-3, 4-6, 7-10
LM-NOM	1-4, 5-10, 11-20, 21-30

Par conséquent, l'ensemble de paires ordonnées

$$A \times B = \{ (a,b) : a \in A, b \in B \}$$

est l'union de deux ensembles disjoints, soit l'ensemble des **concordances**

$$M = \{ (a,b) : a = b, a \in A, b \in B \}$$

et l'ensemble des **non-concordances**

$$U = \{ (a,b) : a \neq b, a \in A, b \in B \}.$$

Les enregistrements relatifs à des éléments de A et B sont désignés respectivement par $\alpha(a)$ et $\beta(b)$. Le **vecteur de comparaison** γ associé aux enregistrements est défini par:

$$\gamma[\alpha(a), \beta(b)] \equiv \{ \gamma^1[\alpha(a), \beta(b)], \gamma^2[\alpha(a), \beta(b)], \dots, \gamma^K[\alpha(a), \beta(b)] \}.$$

Chacun des $\gamma^i, i = 1, \dots, K$, représente une comparaison en particulier. Par exemple, γ^1 peut indiquer une concordance ou une non-concordance pour le sexe. γ^2 peut indiquer qu'il y a concordance entre deux noms de famille et que ceux-ci prennent une valeur précise ou encore qu'il n'y a pas concordance.

Lorsqu'il n'y a aucune confusion possible, nous désignerons la fonction γ sur $A \times B$ par $\gamma(\alpha, \beta)$, $\gamma(a, b)$, ou γ . L'ensemble réalisations possibles de γ est désigné par Γ .

La probabilité conditionnelle de $\gamma(a, b)$, si $(a, b) \in M$, est donnée par

$$m(\gamma) \equiv P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in M\}$$

$$= \sum_{(a,b) \in M} P\{\gamma(\alpha(a), \beta(b))\} \cdot P[(a,b) \mid M].$$

De même, nous désignerons la probabilité conditionnelle de γ si, $(a, b) \in U$ par $n(\gamma)$.

Les règles de décision élaborées dans cet article sont appliquées uniquement aux paires qui renferment généralement les doubles qui sont difficiles à identifier. Les doubles faciles à identifier sont ceux pour lesquels il n'y a généralement pas de problème de concordance des caractères.

Voici un exemple d'un double difficile à identifier:

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Zabinsky Fuel	16 W Sycamore St	Dayton	OH	53315
Zabinsky Cmpny	167 Sycamore St	Springfield	OH	53315

En examinant le second enregistrement, nous constatons que «Zabinsky» et «Sycamore» sont mal écrits, que le terme «Cmpny» est une abréviation inusitée et que le code postal pour Springfield (Ohio), une banlieue de Dayton, est 53315.

2.1.2 Sous-zones particulières comparées

Quatre ensembles de sous-zones particulières font l'objet d'une comparaison dans chaque paire d'enregistrements. Le premier ensemble est celui que l'on peut obtenir par des raisons simples de sous-séquences. On pourrait, par exemple, comparer les caractères des colonnes 1 à 4 de la zone NOM de deux enregistrements.

Dans le tableau 1, on forme la zone LM-NOM en classant les mots de la zone NOM par ordre décroissant de longueur et en scindant les équivalences par un tri alphabétique. On compare ensuite un à un les caractères des sous-zones correspondantes.

Le second ensemble est le résultat des quatre comparaisons des deux plus longs mots de la zone NOM. Là encore, les équivalences sont scindées par un tri alphabétique.

Les deux derniers ensembles sont formés de sous-ensembles des zones RUE et NOM qui sont définis par des logiciels très perfectionnés. Par exemple, ZIPSTAN, qui est un logiciel du Census Bureau (Département du commerce des États-Unis, 1978b), sert à définir des sous-zones correspondantes de la zone RUE. Ces sous-zones sont le numéro civique, les préfixes 1 et 2, le nom de la rue, les suffixes 1 et 2 et l'unité. Les préfixes sont des directions comme «East» ou «North». Les suffixes sont des génériques comme «Street» ou «Road». L'unité représente des identificateurs comme le numéro d'appartement ou le numéro de pièce.

Le module NSKGEN5 du logiciel utilisé dans le Registre des entreprises du Canada (Statistique Canada, 1984, 1982) sert à définir des sous-zones correspondantes de la zone NOM. NSKGEN5 crée trois groupes de mots. Le premier groupe comprend trois abréviations, dont la première correspond au nom de famille si celui-ci est indiqué. Le second groupe est formé de deux mots, dont le premier correspond au nom de famille. Le troisième groupe est constitué en fait d'un seul mot qui est obtenu par l'enchaînement et l'abréviation de mots contenus dans la zone NOM. On trouvera plus de détails à ce sujet dans Winkler (1987) ou dans Statistique Canada (1984, 1982).

2.2 Modèle de Fellegi-Sunter

Le modèle de Fellegi-Sunter utilise une approche théorique décisionnelle qui confirme la validité des principes mis en application par Newcombe (Newcombe et coll., 1959). Pour donner un aperçu modèle, nous le décrivons sous forme de paires ordonnées dans un espace produit. Notre description suit de près celle de Fellegi et Sunter (1969, p. 1184-1187). Soient deux populations A et B dont les éléments seront désignés respectivement par a et b. Nous supposons que certains éléments sont communs aux deux populations.

probabilités des **paramètres d'appariement**. Si l'hypothèse d'indépendance ne se vérifie pas (Winkler 1985c; Kelley 1986), les règles d'appariement fondées sur les probabilités estimées peuvent ne pas être optimales.

Etant donné un intervalle fixe pour les taux d'erreur, de **meilleures** règles d'appariement auront pour effet de réduire l'ensemble des cas indéterminés. Si une règle est fondée sur des paramètres d'appariement qui sont estimés suivant une hypothèse d'indépendance non vérifiée, il peut être possible d'élaborer des méthodes d'ajustement qui permettront de définir de meilleures règles. Pour vérifier si une règle est statistiquement supérieure à une autre, nous utilisons la méthode bootstrap d'Efron (1987; aussi Hall 1988).

Dans les sections qui suivent, nous allons présenter les données de base et des méthodes qui permettent d'appliquer plusieurs règles d'appariement aux listes d'entrepises; nous allons aussi exposer les résultats de l'application de ces règles. Au point de vue de l'application, nous allons utiliser des paires de listes pour lesquelles l'authenticité ou la fausseté des appariements a été vérifiée.

La deuxième section de cet article est divisée en quatre sous-sections. La première contient une description de la base de données et des sous-zones particulières qui sont comparées. La seconde renferme une description sommaire du modèle de Fellegi-Sunter. La troisième met en lumière les hypothèses générales et les méthodes de calcul utilisées. Elle expose aussi en détail les méthodes de calcul qui se rattachent spécifiquement à l'objet de cet article. La quatrième sous-section décrit les méthodes d'évaluation. La méthode d'évaluation fondamentale consiste à suivre l'évolution de la taille de la région des cas indéterminés lorsque divers genres de règles d'appariement sont appliquées suivant un intervalle fixe pour les taux d'erreur. La taille des régions de cas indéterminés est une statistique qui peut dépendre des échantillons ayant servi à uniformiser les règles d'appariement. La distribution de cette statistique est évaluée au moyen de la méthode bootstrap d'Efron (1987, 1982, 1979; aussi Hall 1988). La troisième section de cet article renferme les résultats de l'analyse. Dans la quatrième section, il est question de la robustesse des méthodes de correction de poids, du genre de condition, que représentent les poids corrigés, d'autres genres de comparaisons et de l'application de critères de groupage additionnels. Enfin, nous concluons l'article par un résumé.

2. BASE DE DONNÉES, MODÈLE D'APPARIEMENT, MÉTHODES DE CALCUL ET D'ÉVALUATION

2.1 Base de données

La description de la base de données se fait en deux étapes. Nous allons d'abord décrire les caractéristiques générales de la base, puis énumérer les sous-zones particulières qui font l'objet d'une comparaison.

2.1.1 Description générale

La base de données renferme 54,850 enregistrements, qui représentent chacun une entreprise, et 3,050 doubles, pour un total de 57,900 enregistrements. Une paire d'enregistrements formée d'une entreprise et du double correspondant est désignée comme une concordance; toutes les autres paires sont des non-concordances. La base de données a été construite à l'aide de 11 listes de l'EIA (Energy Information Administration) et de 47 listes fournies par les États et l'industrie; ces listes renfermaient 176,000 enregistrements. Les doubles ont été identifiés par des techniques élémentaires comme le rappel (le numéro de téléphone est parfois indiqué) et le sondage.

Méthodes permettant de tenir compte de l'absence
d'appariement des enregistrements
de Fellegi-Sunter

WILLIAM E. WINKLER¹

RÉSUMÉ

Soit $A \times B$ l'espace produit de deux ensembles A et B , qui est formé de **concordances** (paires dont les éléments représentent la même entité) et de **non-concordances** (paires dont les éléments représentent des entités différentes), en **cas indéterminés** (paires pour lesquelles nous reportons une décision) et en **non-liens** (non-concordances désignées). Suivant un intervalle fixe pour les taux d'erreur, Fellegi et Sunter (1969) ont défini une règle d'appariement optimale, c'est-à-dire une règle qui réduit au minimum l'ensemble des cas indéterminés. L'optimalité dépend de la connaissance de certaines probabilités utilisées dans un rapport de vraisemblance déterminant. En appliquant le modèle d'appariement des enregistrements, on pose souvent une hypothèse d'indépendance qui permet d'estimer les probabilités. Si l'hypothèse n'est pas satisfait, il se peut qu'une méthode d'appariement qui utilise des estimations calculées suivant cette hypothèse ne soit pas optimale. Dans cet article, nous analysons des méthodes qui permettent de modifier les règles d'appariement lorsque l'hypothèse d'indépendance n'est pas valide. À cette fin, nous faisons une analyse empirique de listes d'entreprises pour lesquelles l'authenticité des concordances a été vérifiée. Le nombre de cas indéterminés que produisent les méthodes de calcul habituelles et les méthodes révisées peut varier selon les échantillons. Cette relation est analysée au moyen de méthodes «bootstrap» (Efron 1987).

MOTS CLÉS: Règle de décision; taux d'erreur; (méthode de) la plus grande ascension; bootstrap; saisie-résaisie.

1. INTRODUCTION

Cet article contient une analyse des règles de décision obtenues par suite de l'application du modèle d'appariement des enregistrements de Fellegi-Sunter aux listes d'entreprises. L'analyse vise à comparer une règle obtenue suivant une hypothèse d'indépendance qui est toujours posée en pratique avec des règles qui prévoient des méthodes pour tenir compte de la non-vérification de l'hypothèse d'indépendance. Etant donné deux listes, nous voulons utiliser des identificateurs qui nous permettront de distinguer les paires d'enregistrements dont les éléments se rattachent à la même entité (**concordances**) et celles dont les éléments se rattachent à des entités différentes (**non-concordances**). Nous chercherons donc à définir une règle d'appariement qui nous permettra de diviser l'espace produit de paires en trois groupes: **liens** (concordances désignées), **cas indéterminés** (paires pour lesquelles une décision est reportée) et **non-liens** (non-concordances désignées). Selon un intervalle fixe pour le nombre de concordances et de non-concordances erronées, Fellegi et Sunter (1969, théorème) élaborent une méthode qui, en théorie, réduit au minimum le nombre de cas indéterminés. L'optimalité dépend de la connaissance de certaines probabilités utilisées dans un rapport de vraisemblance fondamental. Dans les applications les plus courantes, on pose une hypothèse d'indépendance qui permet d'estimer les probabilités utilisées dans le rapport de vraisemblance. On appelle ces

¹ William E. Winkler, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233, E.-U.

Techniques d'enquête, juin 1989

103

PLOWDEN, W. (1987). The battles of ideology that ill serve the public. Dans *The Independent*, Londres, le 24 juin 1987.

RATING AND VALUATION ASSOCIATION (1987). Community charge, poll tax: the facts. Rating and Valuation Association, Londres.

REDFERN, P. (1987). A study of the future of the census of population: alternative approaches. Eurostat Theme 3 Series C, Bureau statistique des communautés européennes, Luxembourg.

REMERCIEMENTS

Pour tous les renseignements qui ont servi à la préparation de cette communication, je tiens à remercier les bureaux statistiques de l'Australie, de la Finlande et de la Norvège et, bien entendu, les pays ayant collaboré à l'étude que j'ai réalisée pour la CEE. Quant aux erreurs et lacunes, j'en assume l'entière responsabilité.

BIBLIOGRAPHIE

- AUSTRALIAN HOUSE OF REPRESENTATIVES (1986). The Honorable Neal Blewett MP, à la discussion en deuxième lecture sur l'Australian Card Bill, 1986.
- BOREHAM, J. (1985). Cité dans How Whitehall plays the Numbers Game, *The Times*, Londres, le 30 juillet 1985.
- CHARTERED INSTITUTE OF PUBLIC FINANCE and ACCOUNTANCY (1987). Preparation of a specification of user requirements for the system of community charge in Scotland. CIPFA Services, Londres.
- CITRO, C.F., et COHEN, M.L. (éds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.
- CONSEIL de l'EUROPE (1981). Convention pour la protection des personnes à l'égard du traitement automatique des données à caractère personnel.
- GOVERNMENT STATISTICAL SERVICE (1984). The Government Statistical Service code of practice on the handling of data obtained from statistical inquiries. Cmd 9270, Her Majesty's Stationary Office.
- HEINONEN, R., et LAIHONEN, A. (1987). Some new solutions and methods for census data production: Finnish experiences from the 1985 census. Communication présentée au séminaire de la CEE/CES sur les aspects informatisés des recensements de la population et du logement, Belgique.
- HELDAL, J., SWENSEN, A.R., et THOMSEN, I. (1987). Census Statistics through combined use of surveys and registers? *Statistical Journal of the United Nations Economic Commission for Europe*, 5, 43-51.
- HENNESSY, P. (1987). Why journalists should breach the wall of political secrecy, *The Independent*, Londres, le 1^{er} avril 1987.
- HER MAJESTY'S GOVERNMENT (1986). Paying for local government. Cmd 9714, Her Majesty's Stationery Office.
- HOUSE OF LORDS (1969). The Lord Chancellor, Lord Gardiner, dans Hansard, le 3 décembre 1969. JENSEN, P. (1983). Towards a register-based statistical system - some Danish experience. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.
- JOHANSSON, S. (1987). Statistics based on administrative records as a substitute or a valid alternative to a population census. Communication présentée à la rencontre de l'Institut international de la statistique, Tokyo.
- LAIHONEN, A., et MYRSKYLÄ, P. (1987). Use of registers and administrative records in population censuses in Finland. Communication présentée à la Conférence démographique européenne, Jyväskylä, Finlande.
- MALLET, B. (1917). The organization of registration in its bearing on vital statistics. *Journal of the Royal Statistical Society*, Partie I, 80, 1-24.
- MALLET, B. (1929). Reform of vital statistics: outline of a system of national registration. *Eugenics Review*, 21, 87-94.
- PENRICE, G., REDFERN, P., EVANS, D., WHITEHEAD, F.E., BISHOP, H.E., et RUDOE, W. (1968). Discussion of the Papers on social and medical statistics, *Journal of the Royal Statistical Society*, Sér. A, 131, 26-33.

Tableau 1

Caractéristiques des registres de population utilisés dans 15 pays¹

Registre central	Registres locaux	de population	de coordonner les dossiers administratifs
Numéros de référence	permettant de population	de coordonner les dossiers administratifs	personnels

A. Système complet de registres de population			
Belgique	X		
Danemark	X		
Finlande	X		
Luxembourg	X		
Norvège	X		
Suède	X		
B. Groupe intermédiaire			
France	.		
Pays-Bas	X		
Portugal	.		
Espagne	X		
C. Registres locaux de population seulement			
Rép. fédérale d'Allemagne	X		
Grèce	X		
Italie	X		
D. Aucun registre de population			
Irlande	.		
Royaume-Uni	.		
Nombre de pays possédant cette caractéristique	11	8 +	10

¹ Pour des détails, voir Redfern 1987.

Je crois néanmoins qu'il convient de soulever cette question auprès de statisticiens, et ce pour trois raisons. Premièrement, les statisticiens comprennent à la fois les problèmes techniques et les questions d'intérêt plus général, de sorte qu'ils peuvent indiquer la voie à suivre. Ainsi, au Royaume-Uni, les deux initiatives antérieures concernant les registres de population se sont situées dans un contexte dont la dimension statistique n'était pas exclue (section 5). Deuxièmement, il se peut qu'on donne aux organismes statistiques la responsabilité d'administrer le principal mécanisme de coordination, notamment le registre central de population; c'est le cas de L'INSEE en France et du SSB en Norvège. Troisièmement, les statisticiens gagneraient à disposer de données fiables.

J'espère que les statisticiens feront connaître leur opinion. La question des registres est pleine-ment d'actualité, particulièrement dans les pays aussi «sous-développés» à cet égard que le Royaume-Uni et l'Australie. Les statisticiens au service de l'Etat devraient réfléchir à ce commun-taire sur l'éthique professionnelle formulé à l'occasion de la table ronde sur la méthodologie des recen-sements décennaux qui a lieu en 1984 (Citro et Cohen 1985), mais il est encore très pertinent: «Nous reconnaissons que le climat actuel n'est pas favorable à l'introduction de nouveaux programmes, mais nous pensons que les statisticiens ont la responsabilité de décrire les faits et de recommander les mesures qu'ils jugent raisonnables.»

5) Il y aurait d'autres avantages découlant d'une meilleure vérification des identités. L'ancien registraire général citait comme exemple la possibilité de mieux s'assurer si deux personnes ont le droit de se marier. Le nombre de numéros de référence à donner serait inférieur, et peut-être aussi le nombre de cartes en plastique à avoir sur soi.

6) Les statistiques seraient meilleures, mais, à cet égard, je donnerai des précisions plus loin. Il y a plusieurs réponses à donner aux personnes qui pensent tout de suite Grand Frère et d'énumérer en son nom. En fait, ces personnes n'auraient peut-être pas tort s'il n'existait aucune mesure de protection et si les registres tombaient effectivement dans de mauvaises mains. Mais ces registres peuvent aussi représenter la voie privilégiée vers une société plus équitable. La question qui se pose est la suivante: quel genre de société voulons-nous? Une société qui encourage la fraude, l'évasion fiscale et le crime, ou du moins ferme les yeux sur ces délits? Les ministres australiens ont donné comme exemple un homme qui a été reconnu coupable d'avoir touché, tous les quinze jours, plus de cinquante chèques d'assurance-chômage (Australian House of Representatives 1986). Au Royaume-Uni, en 1987, un avocat, membre du Parlement, a été condamné à la prison parce qu'il avait enfreint les règlements en souscrivant plusieurs fois à des actions, changeant chaque fois de nom, d'adresse et de compte en banque; l'argument invoqué par la défense était qu'il s'agissait d'une pratique courante.

Une autre façon de répondre à l'accusation de totalitarisme consiste à regarder l'utilisation qui est faite des registres de population dans les autres pays. Dans le tableau 1, quinze pays, soit tous les pays de l'Europe occidentale à l'exception de l'Autriche et de la Suisse, sont répartis en quatre groupes selon le genre de système de registres qu'ils emploient. Les six pays du groupe A possèdent le système le plus efficace: leurs dossiers administratifs sont coordonnés au moyen de registres de population. Les quatre pays du groupe B sont dans une situation intermédiaire. Dans les trois pays du groupe C, il existe des registres de population uniquement au niveau local et leur qualité n'est pas toujours bonne. Enfin, l'Irlande et le Royaume-Uni, soit le groupe D, viennent en dernier, avec le système le moins perfectionné. Si le Royaume-Uni choisissait une ligne de conduite rationnelle et réaliste et adoptait le système en vigueur dans le groupe A, il ne se trouverait pas en compagnie de pays totalitaires.

Je dois à présent nuancer ce que j'ai dit au point 6 lorsque j'ai indiqué que les statistiques sont meilleures quant on met en place un système de registres bien coordonnés. La conséquence *directe* d'une telle action est certainement la production de meilleures statistiques, notamment la production de statistiques régionales fiables à intervalles réguliers. Mais si, comme conséquence *indirecte*, les pressions visant à remplacer le recensement classique par un recensement réalisé entièrement à partir de registres se mettent à augmenter de façon irréversible, les avantages sont alors accompagnés d'inconvénients. À la diminution des coûts, du fardeau de réponse imposé au public et des risques de sabotage s'oppose la détérioration possible de la portée et de la qualité des résultats de recensement, notamment dans le domaine économique et dans celui du logement. Il y a alors le danger que les dossiers administratifs rendent de moins en moins bien la complexité et l'évolution des modes de vie actuels, c'est-à-dire ce qu'un recensement classique essaye de mettre en évidence, par exemple l'augmentation du travail à temps partiel et du travail autonome, la croissance du nombre de résidences secondaires et le relâchement des liens familiaux et des liens à l'intérieur des ménages. C'est là que l'expérience des pays scandinaves (section 3) se révèle utile.

Les statisticiens ne sous-estimeront probablement pas l'utilité d'avoir de meilleures statistiques. Cependant, les considérations d'ordre politique et administratif ont plus de poids dans le débat qui entoure les registres de population. Ce sont donc les décideurs, les politiciens et le public qui doivent participer à ce débat. Au Royaume-Uni, celui-ci devrait porter sur le bien-fondé, voire la faisabilité, d'un registre de population qui remplirait une seule fonction, en l'occurrence le prélèvement de la CC, et qui serait totalement indépendant des autres registres existants, par opposition à un registre de population polyvalent qui, par définition, présenterait de nombreux avantages.

6.1 Cartes d'identité

Le projet australien met principalement l'accent sur la carte d'identité comme moyen de vérifier l'identité des gens et non sur le registre ou sur le numéro personnel. Dans certains pays d'Europe, la délivrance d'une carte d'identité est liée à la tenue d'un registre de population, le système en place en Belgique étant l'un des plus perfectionnés. Il est évident que la carte d'identité constitue une mesure supplémentaire de sécurité en autant qu'elle n'est ni contrefaite ni volée. Dans certains pays, comme la France, il n'y a aucun lien entre la carte d'identité et le registre de population.

Dans les pays où l'on n'a jamais utilisé de cartes d'identité en temps de paix, celles-ci sont perçues comme un symbole des régimes autoritaires et comme une atteinte aux libertés civiles. Voilà peut-être une des raisons pour lesquelles le projet australien a suscité une telle opposition de la part du public. Cependant, on peut profiter d'une bonne partie des avantages des registres de population sans avoir recours aux cartes d'identité à condition que les citoyens connaissent leur numéro personnel et le donnent aux autorités lorsqu'ils ont affaire à elles. C'est ce qui se passe au Danemark et en Suède où les cartes d'identité n'existent pas mais où les registres de population sont malgré tout efficaces tant du point de vue administratif que statistique.

Un pays comme le Royaume-Uni ne devrait pas avoir peur de remédier au manque de cohérence de ses dossiers administratifs sous prétexte qu'une critique mal informée confonde le remède nécessaire, en l'occurrence la création d'un registre de population, avec son complément qui, lui, est facultatif, soit les cartes d'identité.

7. CONCLUSION

La création d'un registre de population contenant des adresses à jour et des numéros de référence personnels qu'on retrouverait dans les fichiers administratifs reviendrait tout juste à mettre de l'ordre dans un système boiteux, car même dans le plus beaux des systèmes, il faut bien que le citoyen s'identifie et informe les autorités compétentes de ses changements d'adresse. Il y a malgré tout des gens que l'idée d'un registre de population inquiète parce qu'ils y voient une menace à la liberté et à la vie privée et parce qu'ils craignent qu'un gouvernement autoritaire ou tyrannique n'abuse du pouvoir accru dont l'Etat pourrait ainsi disposer. Pourtant il existe des antidotes auxquels il faudrait avoir recours, notamment des règles efficaces permettant de protéger la confidentialité des données et des lois sur la liberté de l'information.

En revanche, un système de dossiers bien coordonné aurait des avantages politiques qu'on a trop tendance à oublier. Placés par ordre d'importance, les deux premiers seraient, à mon avis:

- 1) Un frein à la fraude, au crime et à l'immigration illégale.
- 2) Une société plus juste dont les devoirs seraient mieux partagés et dont les privilèges seraient uniquement à ceux qui y ont droit. Autrement dit, la liberté ne devrait pas signifier le loisir de frauder le reste de la collectivité.

Plus loin, j'ajouterais:

- 3) Les économies financières que l'Etat pourrait réaliser. Si les dossiers étaient plus précis, les frais administratifs seraient inférieurs, il serait possible de prélever davantage de taxes et le montant des prestations versées à tort serait réduit, comme le montrent les chiffres australiens (Section 6).
- 4) Le gouvernement aurait plus de choix quand aux moyens à sa disposition pour mettre en oeuvre ses politiques. Si, par exemple, il existait déjà un registre de population au Royaume-Uni, le gouvernement n'aurait pas à en créer un *spécial* pour son projet de charge collective et il pourrait surveiller l'entrée des immigrants en ajoutant aux contrôles exercés aux aéroports et aux ports de mer un contrôle à partir de l'adresse du domicile.

L' *Australia Card Bill* de 1986 (projet de loi concernant la carte australienne) a été approuvé par la Chambre des députés mais non par le Sénat, dont le parti au pouvoir ne détient pas la majorité des sièges. Apparemment, c'est en partie à cause de ce rejet que l'élection de 1987 a eu lieu. Après la victoire du parti au pouvoir, le Parlement devait être saisi du projet de loi de nouveau, mais, celui-ci a été retiré à cause d'une sérieuse imperfection légale. Je pense qu'il est quand même utile de décrire ici les dispositions qu'il contenait.

Le registre de l'AC serait un registre central de population. Cependant, il serait moins complet que ceux des pays scandinaves pour deux raisons principales:

- 1) En vertu du projet de loi, les citoyens ne seraient pas obligés de signaler leurs changements d'adresse aux responsables du registre. Si j'ai bien compris, on espérait que la plupart des changements d'adresse seraient communiqués à au moins un des organismes publics liés au projet, lequel en ferait part au service chargé de la mise à jour du registre de l'AC.
- 2) Le projet de carte australienne ne serait pas aussi polyvalent que plusieurs des registres de population européens. À cause des inquitudes du public relativement à la protection de la confidentialité des renseignements personnels et au couplage incontrôlé des données, seuls les organismes publics qui s'occupent des impôts, de la sécurité sociale et de l'assurance-maladie auraient accès au registre de l'AC, et encore, uniquement pour vérifier les identités.

Le projet de loi précisait dans quelles situations on pouvait demander à quelqu'un de montrer sa carte, ce serait notamment à l'occasion d'un large éventail d'opérations financières et au moment de commencer un nouvel emploi, de se faire soigner à l'hôpital et de demander des prestations ou des services offerts par l'assurance-maladie ou la sécurité sociale. Il serait illégal de demander à quelqu'un de montrer sa carte dans d'autres circonstances.

Comme mesure supplémentaire de protection des renseignements personnels, le projet de loi prévoyait la création d'un organisme chargé de la protection de la confidentialité des données. Toutefois, le gouvernement est d'avis qu'il faut trouver le juste milieu entre la protection de la vie privée et les pertes que la fraude fiscale fait subir à l'État. Il estime que le projet d'AC coûterait \$0.8 milliard en dix ans, mais que cette dépense serait plusieurs fois compensée par les montants de \$4.1 milliards et de \$1.4 milliards que le fisc et la sécurité sociale pourraient récupérer respectivement, permettant à l'État de réaliser pendant cette période des économies nettes de l'ordre de \$4.7 milliards (Australian House of Representatives, 1986).

Les remarques formulées par le ministre de la Santé au Parlement sont révélatrices des buts poursuivis par les ministres et de l'engagement politique sans équivoque pris par le gouvernement:

«Je tiens à saisir le Parlement aujourd'hui . . . d'une réforme qui s'est longuement faite attendre et qui vise à apporter justice et équité à tous les Australiens.»

Il ne fait aucun doute que la carte australienne permettra de contrôler la fraude fiscale; il ne fait aucun doute qu'elle contribuera à préserver l'intégrité de notre régime de sécurité sociale; il ne fait aucun doute qu'elle s'avèrera une arme efficace contre l'immigration illégale; il ne fait aucun doute qu'en permettant de remonter la piste empruntée par l'argent, elle sera un instrument précieux de lutte contre la criminalité des entreprises et le crime organisé.»

«Il faut protéger les citoyens contre l'intrusion du gouvernement dans leur vie privée, c'est un principe absolu. Mais il faut également protéger les citoyens contre ceux qui se cachent cyniquement derrière le droit à la confidentialité des renseignements personnels pour assumer une fausse identité et frauder la collectivité.»

«Notre pays établira un système d'identification avant la fin du siècle, c'est inévitable.»

Bien que le projet de loi concernant la carte australienne ait été retiré, ce n'est pas encore la fin de cette histoire, car le gouvernement continue à chercher d'autres moyens qui lui permettront de supprimer toute possibilité de fraude à l'impôt et à la sécurité sociale.

Au Royaume-Uni, le principal obstacle à l'instauration d'un registre de population est la résistance traditionnelle du public à toute mesure gouvernementale perçue comme étant autoritaire ou bureaucratique. On peut s'attendre que le lobby de la protection de la vie privée mène l'opposition contre toute nouvelle obligation, pour le public, de déclarer des renseignements, tout ajout au nombre de données à caractère personnel que détient déjà l'Etat et tout projet de couplage de données. L'opposition ne tient pas compte des coûts et des injustices qu'entraîne la gestion inefficace des données. Elle ne tient pas compte non plus des freins que les lois sur la protection de la confidentialité des données et la liberté de l'information peuvent mettre, si elles sont appliquées, à la mauvaise utilisation des données à caractère personnel, ou alors elle sous-estime ces freins. La perception que le public a du gouvernement en place a renforcé ces dernières années les craintes qu'il éprouve à l'idée de fournir davantage de renseignements personnels à l'Etat: le gouvernement britannique est perçu comme ayant l'obsession du secret et comme cherchant à garder tout le pouvoir entre ses mains. Ainsi, non seulement n'y a-t-il pas de loi sur la liberté de l'information au Royaume-Uni, mais en plus les renseignements concernant l'Etat ont en principe été protégés par une loi d'application générale, la *Official Secrets Act de 1911* (loi sur les secrets officiels). Peter Hennessy, rédacteur de *Con-temporary Record* affirme que les gouvernements britanniques ont, de tous les gouvernements occidentaux, adopté les mesures les plus rigoureuses pour protéger le secret des dossiers administratifs (Hennessy 1987). D'ailleurs, à cause d'événements récents, la nécessité pour les services secrets de justifier leurs actions a fait l'objet de discussions visant à mieux définir cette responsabilité. S'exprimant sur l'ensemble des activités gouvernementales, William Plowden, directeur général du Royal Institute of Public Administration, a dit qu'un gouvernement britannique contemporain, appuyé par une bonne majorité à la Chambre des communes, peu menacé par des commissions parlementaires qui aboient mais ne mordent pas, solidement protégé par la loi sur les secrets officiels, est, de tous les pouvoirs exécutifs des pays industrialisés, un de ceux qui sont le moins tenus de rendre compte de leurs décisions (Plowden 1987).

Le public se méfie donc de tout nouveau projet visant à créer un registre de population. Et, ainsi que nous l'avons vu, le gouvernement actuel a exprimé son opposition à l'idée d'un registre complet: comme celui des Etats-Unis, il a montré qu'il était décidé à lutter contre toute forme d'ingérence de l'Etat dans la vie des citoyens. Un de ses principaux objectifs est de réduire la taille et l'influence du secteur public, et il accorde parfois plus d'importance à ce principe qu'à celui de la rentabilité. On voit donc que les préoccupations du public relativement à la protection de la vie privée, l'idéologie politique et le manque de ressources sont trois facteurs qui se conjuguent pour empêcher la création d'un registre complet, lequel permettrait pourtant de réaliser des économies considérables et de rendre la société plus équitable. A vrai dire, les faits concernant ces questions n'ont pas été présentés de façon équilibrée, et il n'y a pas eu de débat public à ce sujet, depuis une cinquantaine d'années.

6. L'INITIATIVE AUSTRALIENNE: LES CARTES D'IDENTITE

Je ne connais pas bien le tempérament australien ni la situation politique dans ce pays, mais je suppose que l'opposition à un gouvernement bureaucratique est aussi forte là-bas qu'au Royaume-Uni. Malgré cela, le gouvernement australien a présenté un projet de loi visant la délivrance d'une carte d'identité à tous les citoyens, l'AC (Australia Card/carte australienne). Les raisons sont purement administratives: il s'agit de réduire l'évasion fiscale, la fraude en matière de la sécurité sociale et l'immigration illégale. La carte australienne porterait le nom, la photo, la signature et le numéro (numéro de référence personnel propre à cette carte d'identité de chaque personne), mais non son adresse. La carte australienne serait rattachée à un registre qui contiendrait les adresses et les dates de naissance, et auquel n'auraient accès que certains ministères ou services publics.

La création des registres de la CC représente peut-être une occasion manquée de constituer un registre de population efficace. Mais en fait, le projet de CC n'est pas le prétexte idéal pour cela. En effet, pour être efficace, le registre de population doit servir à plusieurs fins, et plus il y en a, mieux c'est. Il ne devrait pas avoir une seule fonction, surtout lorsque celle-ci est de permettre de prélever une taxe que beaucoup trouvent lourde et à laquelle un grand nombre de gens chercheront à se dérober. De plus, la CC est une mesure politique controversée parce qu'elle n'a pas le même effet sur les divers segments de la société: elle entraînera de façon générale un transfert des ressources des pauvres aux riches.

Il y a donc plusieurs raisons de mettre en doute l'efficacité opérationnelle des registres qui doivent être créés dans le cadre du projet de CC: le but unique et controversé des registres; le fait que toute la population n'y sera pas représentée (l'omission de certains groupes); l'absence de registre central permettant de coordonner les registres locaux; et enfin, le fait que ce sont le nom et (en Ecosse) la date de naissance qui serviront d'identificateur, plutôt qu'un numéro personnel permanent. Les autorités locales ont exprimé leurs réserves à l'égard de ce projet en décrivant les problèmes auxquels elles devront faire face au moment de constituer les registres (Rating and Valuation Association 1987). Il semble bien que le gouvernement se soit engagé à introduire une nouvelle loi fiscale sans avoir pensé à tous les aspects pratiques de son application. Il y a un autre aspect du projet de CC qui est inquiétant, et c'est l'effet qu'il aura sur la réaction du public au recensement de la population de 1991. Parmi ceux qui voudront se dérober à la CC, un bon nombre essaieront aussi de se soustraire au recensement, malgré les efforts que feront les responsables de ce dernier pour les convaincre que les données recueillies ne seront pas communiquées à d'autres organismes. Par contre, si le questionnaire dit de façon trop explicite: «LES RENSEIGNEMENTS QUE VOUS DÉCLAREREZ NE SERONT PAS COMMUNIQUÉS AUX ORGANISMES CHARGÉS D'ADMINISTRER LES IMPÔTS, LA SÉCURITÉ SOCIALE, LES CHARGES COLLECTIVES, ...», les responsables du recensement ne risquent-ils pas de donner l'impression qu'ils ne condamneront pas, ou même qu'ils encouragent, l'évasion et la fraude?

5.6 La situation au Royaume-Uni

Indépendamment de la CC, le climat actuel au Royaume-Uni est plutôt hostile à l'idée d'un registre de population. Mais on peut mentionner deux aspects positifs. Premièrement, la *Data Protection Act de 1984* (loi sur la protection de la confidentialité des données) a introduit des mesures de protection touchant les renseignements personnels stockés dans les ordinateurs qui sont semblables à celles prévues par la Convention de 1981 du Conseil de l'Europe (Conseil de l'Europe 1981). En fait, le principal but visé par le gouvernement lorsqu'il a adopté la loi de 1984 était un but commercial: il s'agissait de montrer aux autres pays susceptibles d'envoyer leurs données au Royaume-Uni pour les faire traiter que ces données y seraient en sécurité. La protection de la vie privée était un but secondaire. Deuxièmement, le GSS, qui serait chargé de voir à certains aspects du fonctionnement des registres de population, a toujours été irréprochable en ce qui a trait à la protection de la confidentialité des données; il a d'ailleurs publié un code de conduite à ce sujet (Government Statistical Service 1984). Le fait que le GSS soit décentralisé a également contribué à établir la réputation d'intégrité que cet organisme s'est méritée, l'échange de données, même à des fins statistiques, étant empêché par des obstacles tant juridiques qu'administratifs. Il faudrait d'ailleurs retirer ces obstacles pour bénéficier des retombées statistiques qu'un registre de population pourrait avoir.

Pour ce qui est des aspects négatifs, la dépendance du GSS à l'égard du gouvernement central contraste avec l'autonomie relative dont jouissent notamment les organismes statistiques du Danemark et des Pays-Bas. Cette dépendance pourrait ébranler la confiance que le public accorde au GSS relativement au traitement des données. L'impression que peut donner le GSS d'être la marionnette du gouvernement central a été renforcée par les conclusions du rapport Kayner au début des années 80. À la suite de ce rapport, on a demandé au GSS d'accorder plus d'importance aux besoins du gouvernement central, et ce aux dépens des autorités locales, des entreprises, du milieu universitaire et du grand public.

changement d'adresse. On ne créera pas pour autant de registre de population à proprement parler. Le gouvernement s'y oppose absolument.

La nouvelle obligation de déclarer tout changement d'adresse, qui constitue une dérogation de taille à la tradition britannique en temps de paix, découle de la décision du gouvernement de changer la façon dont les taxes locales sont perçues. Par le passé, le montant de la taxe locale imposée aux occupants d'une propriété était déterminé par la valeur locative de la propriété. Cette taxe va être remplacée par la CC (*Community Charge*/charge foncière collective), qui est une taxe uniforme que devra payer toute personne âgée de 18 ans et plus habitant un logement. Pour administrer la nouvelle taxe, il faudra tenir un registre local des adresses indiquant le nom des personnes âgées de 18 ans et plus qui y habitent. Le responsable des inscriptions pourra faire enquête et adresser des demandes de renseignements aux autorités locales, aux commissions du logement et au bureau des élections, mais c'est à l'individu qu'incombent la responsabilité de l'informer des changements à apporter au registre. La loi visant à introduire ce nouveau système a été adoptée en Ecosse où elle est en vigueur depuis avril 1989, et sera en vigueur en Angleterre et au pays de Galles à compter d'avril 1990.

Cependant, les registres de la CC seront des instruments primitifs comparativement aux registres de population des pays scandinaves et des pays du Benelux, pour les raisons suivantes :

- 1) Les registres de la CC n'incluront pas tout le monde; en seront exclus notamment les moins de 18 ans et les résidents d'une pension ou d'un établissement spécialisé.
- 2) Les registres (où seront consignés le nom, l'adresse et, en Ecosse, la date de naissance des individus) seront tenus localement et la marche à suivre à leur égard ne sera pas pleinement uniformisée. Ainsi, il n'y aura pas de registre central qui imposerait une même façon de décrire l'identité de chaque personne inscrite et assurerait la coordination des registres locaux (afin de faciliter les transferts d'une autorité compétente à une autre, par exemple).
- 3) Bien que la loi ne prévoie pas explicitement l'utilisation d'un numéro de référence personnel dans les registres, il a été recommandé dans un rapport que les autorités locales écossoises créent un nouveau numéro, et un algorithme pouvant servir à l'établissement de ce numéro au moyen du nom et de la date de naissance a été proposé (Chartered Institute of Public Finance and Accountancy 1987). Mais la recommandation n'a pas été exécutée.

- 4) La loi précise qui peut avoir accès à quelles parties du registre. À part les autorités locales qui y auront accès pour pouvoir administrer la CC, un particulier pourra examiner les renseignements qui le concernent, le public pourra examiner les listes d'adresses et le nom des personnes correspondant à ces adresses («pas en vue de s'assurer si une personne habite bien à une certaine adresse») et le directeur général des élections aura accès aux registres pour les besoins liés à ses fonctions. Personne d'autre n'y aura accès.

Le rejet par le gouvernement d'un registre de population qui permettrait de coordonner les dossiers administratifs est justifiée dans le livre vert sur le projet de CC (Her Majesty's Government 1986). Les auteurs de ce document citent le cas de pays qui «ont fusionné leur divers registres et s'en servent centralement à plusieurs fins administratives différentes». Ils ajoutent : «La tradition britannique est différente. Les registres sont tenus séparément à des fins différentes par les organismes qui en ont besoin dans un but particulier . . . Il n'y aura pas de registre national.» La comparaison qui est faite entre la pratique observée dans les autres pays et celle proposée pour le Royaume-Uni est fautive, car, dans les autres pays, les organismes tiennent des registres *distincts* mais s'adressent au service chargé du registre central pour identifier la personne à laquelle ils ont affaire. À mon avis, la déclaration «il n'y aura pas de registre national» découle d'un axiome politique et non d'une analyse rationnelle.

- 1) Une forte proportion des données destinées au NHSCR ne sont pas assorties du numéro d'identification personnel. Du fait que le nom et la date de naissance peuvent difficilement servir d'identificateur à eux seuls, certaines données ne peuvent pas être intégrées aux dossiers existants du NHSCR. C'est notamment ce qui arrive avec 1 à 2 % des avis de décès. C'est en grande partie pour cette raison et parce qu'on arrive pas à éliminer tous les immigrants du registre que le nombre de personnes inscrites dans le registre est supérieur à ce qu'il devrait être. Ces cas, dont la proportion est estimée actuellement à environ 5%, devraient bientôt être moins nombreux, lorsque le registre sera informatisé.
- 2) Les adresses figurent au complet dans les registres locaux et sont représentées par un code régional dans le NHSCR. Cependant, dans la plupart des cas, les changements d'adresse ne sont effectués que lorsqu'un personne s'inscrit auprès d'un nouveau médecin, ce qui peut survenir des années après qu'elle a déménagé.

5.3 L'éventail des registres au Royaume-Uni

Comme dans tout autre pays industrialisé, les autorités du Royaume-Uni tiennent une vaste gamme de registres contenant des renseignements personnels. Les principaux registres concernent les actes d'état civil (naissances, décès, mariages et divorces), l'immigration et la naturalisation, le service de santé national, la sécurité sociale (tant cotisants que bénéficiaires, comme les chômeurs, les retraités et les enfants), les impôts des particuliers, les passeports, les listes électorales, la possession d'une voiture et le permis de conduire. Mais ces registres sont tenus indépendamment les uns des autres par des organismes distincts qui ont chacun leur système d'attribution de numéro personnel. Il y a une exception, et c'est l'entente conclue dans le cadre du système de retenue à la source pour la collecte conjointe des impôts sur le revenu et des cotisations versées par les employés à la sécurité sociale. Un seul numéro personnel est utilisé, le numéro d'assurance sociale. À part ça, les systèmes de dossiers ne sont coordonnés d'aucune façon; le contenu des dossiers n'est pas uniformisé et il n'y a pas un numéro personnel qui est utilisé de façon générale. Les renseignements relatifs à l'identité d'une personne, habituellement le nom et la date de naissance, peuvent varier d'un registre à un autre et même à l'intérieur d'un même registre. Par conséquent, il arrive que les mêmes personnes figurent plus d'une fois dans les registres, et il n'est pas certain qu'on puisse combiner les renseignements que ces derniers contiennent à des fins statistiques. Si on pouvait le faire, ce serait très coûteux. Les renseignements concernant les adresses sont, quant à eux, encore moins cohérents. Enfin, il n'existe pas de mécanisme permettant de mettre à jour simultanément tous les dossiers visés par certains changements, notamment les changements d'adresse, le changement de nom après le mariage ou même les décès. Comme l'a si bien dit Sir John Boreham lorsqu'il dirigeait le GSS (Government Statistical Service): «les renseignements ne sont jamais réunis convenablement . . . tout le système est plutôt boiteux» (Boreham 1985).

5.4 Étude réalisée dans les années 60

Le système actuel de registres indépendants ne peut pas être administré de façon efficace. Et du point de vue statistique, il souffre d'un double handicap: les adresses ne sont pas à jour et on ne peut pas appairer les dossiers. Le GSS s'est donc mis à la recherche d'une solution vers la fin des années 60. Il a étudié la possibilité de remplacer les divers types de numéros personnels par un seul numéro qui serait consigné dans le registre central, lequel contiendrait éventuellement la dernière adresse des personnes inscrites (Penrice et coll. 1968). Toutefois, les ministres ont décidé que ces idées étaient politiquement inacceptables, et ils ont mis fin à l'étude (House of Lords 1969).

5.5 Les registres pour la nouvelle charge foncière collective

Il semble qu'un des plus grands obstacles à la création d'un registre de population en Angleterre a été surmonté puisque maintenant on oblige les citoyens à déclarer tout

Les quatre pays semblent prêts à sacrifier en partie la qualité des résultats si cela permet de réduire les coûts et d'alléger le fardeau de réponse imposé à la population. Mais ils n'adoptent pas tous la même approche. C'est le Danemark qui, en abandonnant le questionnaire du recensement, est allé le plus loin. Comme la qualité des données provenant de certains registres était mise en doute, particulièrement en ce qui a trait aux questions de nature économique, la Finlande et la Suède ont conservé un questionnaire restreint pour le recensement de 1985 et elles ont combiné les réponses obtenues aux données démographiques et autres tirées des registres. Cependant, il est possible qu'en 1990 le recensement de la Finlande soit entièrement réalisé à partir de registres. En Norvège, où il n'y a pas de recensement quinquennal, il est prévu qu'au recensement de 1990 on conserve le questionnaire au moins pour les questions d'ordre économique, mais qu'on ne l'envoie qu'à un échantillon de 10% de la population, dans le but de réduire les coûts. Les données de nature économique qui se trouvent dans les registres pourront être corrigées d'après les résultats obtenus auprès des échantillons afin de devenir compatibles avec les définitions statistiques retenues. Johansson (1987) donne un compte rendu très utile de l'expérience suédoise relative à l'utilisation des registres comme source de données pour le recensement.

4. POSSIBILITÉ DE RÉALISER DES RECENSEMENTS À PARTIR DE REGISTRES DANS LES AUTRES PAYS

Les deux principaux facteurs qui ont poussé les pays scandinaves à réaliser leurs recensements à partir de registres, soit la nécessité de réduire les coûts et d'alléger le fardeau de réponse, ne sont pas pour autant négligés ailleurs. Ainsi, ils ont eu pour effet de freiner, et parfois de renverser, la tendance observée avant 1980 selon laquelle les questionnaires de recensement devenaient de plus en plus longs.

Un nouvel élément troublant, l'opposition du public, a perturbé le recensement dans deux pays. Aux Pays-Bas, le projet de réaliser un recensement en 1981 a été abandonné. En République fédérale d'Allemagne, le recensement prévu pour 1983 dû être reporté jusqu'en 1987 à cause des exigences plus strictes que la cour constitutionnelle a imposées en matière de confidentialité, et malgré cela tout le monde n'y a pas participé. Aucun pays n'est à l'abri de ce genre de contestation. Cependant, un recensement réalisé à partir de registres risque moins d'être saboté dans la mesure où il n'a pas besoin d'être complet par un questionnaire. C'est qu'un tel recensement ne crée pas une occasion (le jour du recensement) où tout le monde a un questionnaire à remplir et où les protestations d'une minorité peuvent s'étendre jusqu'à prendre la forme d'une opposition massive.

Si un recensement réalisé à partir de registres coûte tellement moins cher et réduit à la fois le fardeau de réponse et les risques de sabotage, pourquoi si peu de pays considèrent-ils qu'il s'agit d'un choix praticable? Pour trois raisons. Premièrement, dans certains domaines, particulièrement le domaine économique, les données administratives sont parfois de qualité inférieure aux données recueillies au moyen d'un questionnaire, tandis que dans d'autres domaines, il n'existe pas de données administratives. Les pays scandinaves sont conscients de ces problèmes, et c'est pourquoi certains utilisent encore un questionnaire et combinent les réponses ainsi recueillies avec les données tirées des registres (section 3.5).

Deuxièmement, un grand nombre de pays ne possèdent pas un système de données du genre de celui décrit à la section 2. Par exemple, certains pays comme la République fédérale d'Allemagne, la Grèce et l'Italie ont des registres de population locaux mais pas de registre central de population. Ou alors les registres de population ne sont pas à jour, et, dans certains pays comme l'Italie et l'Espagne, on compte même sur le dénombrement de la population effectué dans le cadre d'un recensement classique pour leur mise à jour. À part les pays scandinaves, il y a les pays du Benelux qui ont, ou auront probablement bientôt, l'infrastructure nécessaire à la tenue d'un recensement réalisé à partir de registres.

Après le recensement de 1980, on avait fait une étude portant sur les mesures qu'il faudrait prendre pour que le recensement de 1985 soit entièrement réalisé à partir de registres. Parmi ces mesures, il y avait les suivantes:

- 1) Utiliser les données sur la profession tirées des formules dans lesquelles les personnes occupées déclarent tout changement de revenu au bureau de l'assurance nationale.
- 2) Créer un registre de la composition des ménages qu'on mettrait à jour en recueillant des renseignements à l'occasion des déménagements.
- 3) Créer un registre des immeubles qui contiendrait des données sur les unités de logement et qui serait mis à jour par les municipalités.
- 4) Créer un registre des études terminées qui serait mis à jour à l'aide des renseignements fournis par les établissements d'enseignement sur les diplômés décernés.

Comme on l'a vu plus haut, on a conservé le questionnaire au recensement de 1985 principale-ment parce qu'on avait des doutes au sujet de la qualité des renseignements qui pourraient être tirés des registres relativement à la profession, à la composition des ménages et au logement. De tous les nouveaux registres proposés, seul celui sur les études terminées est en voie d'être constitué. Cependant, un comité étudie actuellement la possibilité d'indiquer l'unité de loge-ment avec l'adresse dans les registres de population, ce qui est essentiel si l'on veut procéder au couplage des registres de population et des registres des logements.

Une commission parlementaire est en train de se pencher sur le recensement de 1985, parti-culièrement sur les aspects relatifs à la vie privée et à la confidentialité. Les conclusions aux-quelles elle arrivera contribueront à déterminer la forme que prendra le recensement de 1990.

3.5 Les recensements en Scandinavie: résumé

L'évolution du recensement dans les quatre pays scandinaves se fait selon des voies diffé-rentes, mais il y a de nombreux points communs:

- 1) Tous prennent comme point de départ des registres de population exacts permettant de produire régulièrement des statistiques régionales fiables.
- 2) Tous souhaitent maximiser l'utilisation des renseignements contenus dans les autres registres et minimiser le fardeau de réponse imposé au public. Tous s'efforcent de limiter ou de réduire les coûts.

- 3) Tous reconnaissent que les renseignements contenus dans les registres, particulièrement les renseignements de nature économique, posent des problèmes en ce qui a trait à défini-tions, à la qualité et aux périodes de référence. On est en train d'augmenter le nombre de renseignements que les employeurs doivent déclarer, notamment en ce qui concerne le lieu de travail de chaque employé et donc la branche d'activité. Cependant, le fait qu'on demande des renseignements supplémentaires à des fins purement statistiques est mal accueilli, ce qui risque d'avoir pour conséquence une baisse de la qualité des données. Les données sur la profession qu'on trouve dans les registres ne sont en général pas fiables, et il y a des sujets sur lesquels on ne trouve rien du tout, notamment les moyens de transport utilisés pour se rendre au travail.

- 4) Des registres des immeubles et des logements ont été créés, ou du moins sont à l'état de projet. Il est difficile de tenir certains registres à jour, que ce soit en ayant recours aux renseignements dont disposent les municipalités ou en recueillant les renseignements nécessaires directement auprès des propriétaires. Dans certains pays, il faudrait améliorer les registres en identifiant chaque unité de logement de manière à pouvoir faire le lien avec les adresses inscrites dans les registres de population. Il y a un autre problème: comment obtenir des données sur la composition des ménages si, comme en Suède, le ménage n'est pas défini comme étant constitué de tous les occupants de l'unité familiale.

on a envoyé à chaque personne âgée de 16 ans et plus un questionnaire qu'elle devait retourner par la poste et qui comportait des questions de nature économique et des questions sur les études faites à l'étranger, sur le pays de naissance, sur la religion et sur le logement. Toutes les personnes faisant partie d'un même ménage devaient renvoyer leur formule, ainsi qu'une formule sur le logement, dans la même enveloppe, établissant ainsi la composition du ménage aux fins du recensement.

Il y a plusieurs raisons pour lesquelles il ne sera pas possible, en 1990, de faire un recensement réalisé entièrement à partir de registres. Premièrement, les données des registres relatives à certaines variables importantes du recensement ne sont pas compatibles avec les définitions statistiques retenues ou ne sont pas d'assez bonne qualité pour le recensement (c'est le cas notamment des données relatives à la branche d'activité) ou encore les registres ne contiennent pas de données se rapportant à certaines variables (par exemple la profession). Deuxièmement, il est peu probable que le registre des propriétés foncières, des adresses et des immeubles (le «GAB») qu'on a commencé à constituer en 1983 soit assez avancé en 1990 pour qu'on puisse en tirer des données sur le logement. Troisièmement, comme c'est l'adresse qui constitue le lien entre le GAB et le registre de population, il n'est pas possible de déterminer la composition du ménage ni d'associer caractéristiques du logement et caractéristiques personnelles lorsque plusieurs unités de logement portent la même adresse.

Pour le recensement de 1990, on ira encore chercher dans les registres les données démographiques de base, celles sur le revenu et celles sur les études terminées (autres que les études faites à l'étranger). Par ailleurs, ont est en train de mettre au point une méthode qui permettra de modifier les données des registres concernant la plupart des variables économiques en fonction des résultats d'un sondage mené auprès d'un échantillon de 10% des personnes âgées de 16 ans et plus (100% dans les municipalités de moins de 6,000 habitants), afin qu'elles soient compatibles avec les définitions statistiques retenues. Pour corriger les données des registres se rapportant à une sous-population, on se servira, d'une part, des données recueillies auprès d'un échantillon de cette sous-population et, d'autre part, des données recueillies auprès d'un échantillon d'une population dont l'effectif est plus grand, ce qui aura pour effet d'éliminer en partie le biais dont sont entachées les données des registres. Le sondage sera la seule source de données pour les sujets sur lesquels il n'existe rien dans les registres, c'est-à-dire la profession et probablement aussi le logement et la composition des ménages.

On estime le coût de cette approche, soit l'utilisation des registres et la tenue d'un sondage auprès d'un échantillon de 10% de la population visée, à 60% du coût d'un recensement comme celui de 1980. Le prix à payer sera, premièrement, la variabilité d'échantillonnage, qui sera à son plus fort pour les sujets sur lesquels il n'existe aucune donnée dans les registres, et, deuxièmement, un certain biais dans le cas des données contenues dans les registres mais dont la qualité n'est pas celle qu'on recherche pour un recensement (Heldai et coll. 1987).

3.4 Suède

En ce qui concerne le recensement de la Suède, la situation s'est inversée au cours des vingt dernières années: en 1970, la plupart des données provenaient des questionnaires et un petit nombre de registres; en 1985, c'était le contraire. Cette année-là, dans le questionnaire envoyé et retourné par la poste, on demandait à chaque personne âgée de 16 ans et plus (ou couple marié) d'indiquer uniquement: (1) si elle était active au cours d'une semaine précise et, le cas échéant, quel métier ou profession elle exerçait; (2) la composition du ménage, soit la liste des adultes habitant dans le même logement; (3) des renseignements sur le logement. Il a été possible d'omettre les questions sur le nom de l'entreprise pour laquelle la personne travaillait, le lieu de travail et la branche d'activité, questions qui avaient été posées au recensement précédent, parce qu'on a demandé aux employeurs d'ajouter un renseignement à leur déclaration d'impôt annuelle, en l'occurrence le lieu de travail de chaque employé. Par contre, les employeurs se sont opposés à indiquer également dans leur déclaration d'impôt le nombre d'heures travaillées, et ce sujet a donc été abandonné au recensement de 1985.

3.2 Finlande

d'obtenir des données fiables sur la profession, parce que ce sujet revêt peu d'intérêt du point de vue administratif; la principale source d'information est la déclaration d'impôt annuelle. Malgré ces problèmes, Danmarks Statistik considère que les recensements réalisés à partir de registres constituent une réalité bien ancrée au Danemark parce qu'ils permettent de réduire les coûts et d'alléger le fardeau de réponse imposé au public (Jensen 1983).

Les recensements réalisés à partir de registres ont une longue histoire en Finlande. Au 17^e siècle, tous les membres d'une paroisse âgés de plus de 12 ans étaient inscrits dans les registres paroissiaux, et en 1749 on a compilé les données se rapportant à la population entière et on les a analysées selon l'âge, le sexe, l'état matrimonial et la classe sociale. Étaient-ce là un des premiers recensements jamais réalisés à partir de registres? Les recensements suivants ont été effectués de la même manière, mais en 1950 et en 1960 on a adopté la méthode classique qui consiste à recueillir des renseignements au moyen d'un questionnaire. Cependant, depuis le recensement de 1970, on tire des registres un éventail de plus en plus large de données. Les questions posées au recensement quinquennal de 1985 appartenaient uniquement au domaine économique; genre d'activité (le cas échéant) et situation professionnelle, employeur et lieu de travail, profession et nombre de mois travaillés pendant l'année précédente. Les données sur le logement ont été tirées du registre des immeubles et des logements créé à partir des données du recensement de 1980 et mis à jour à l'aide des renseignements fournis par les municipalités.

Le recensement de 1985 a été conçu de manière à coûter un peu moins que l'équivalent d'un dollar américain par personne, soit le *quart* du coût du recensement de 1980 en termes réels, tout en comportant la même gamme de variables. Parmi les facteurs qui ont permis de réaliser cet exploit, mentionnons: les questionnaires envoyés par la poste sur lesquels étaient pré-imprimées les données sur le lieu de travail (d'après les renseignements recueillis au recensement de 1980) et sur la profession (d'après les renseignements inscrits dans le registre central de population) que les recenseurs pouvaient corriger au besoin; le retour des questionnaires par la poste directement au bureau central, sans passer par une organisation locale; un seul rappel, et aucun suivi dans le cas des formulaires qui n'avaient pas été renvoyés ou qui étaient incomplets (3,7%); et l'imputation des données manquantes en ayant recours, dans le mesure du possible, à divers registres, dont les dossiers des régimes de retraite pour ce qui est des employés du secteur privé. Le taux de réponse au questionnaire était de 97,4% et, après l'imputation des données manquantes, le taux de couverture final était de 98,6%. Le faible coût du recensement est également attribuable au fait que le fardeau financier et administratif a été assumé en partie par les services chargés d'administrer les registres, dont la vérification annuelle sur le terrain des registres de population par l'envoi de formulaires à chaque ménage ou logement et la vérification quinquennale du registre des immeubles et des logements par l'envoi de formulaires aux propriétaires et aux occupants.

La comparaison des données tirées du recensement de 1980 et des données se rapportant aux variables économiques tirées des registres a donné des résultats considérés comme encourageants. C'est pour cette raison et aussi grâce aux méthodes d'imputation des caractéristiques économiques des non-répondants élaborées à l'occasion du recensement de 1985 qu'il y a de bonnes chances que le recensement de la Finlande soit entièrement réalisé à partir de registres en 1990. Pour combler la seule lacune des données tirées de registres, les employeurs possédant plus d'un établissement devront à l'avenir déclarer le lieu de travail de chaque employé (Laiho et Myrskylä 1987; Heinonen et Laiho 1987).

3.3 Norvège

Le recensement de la Norvège de 1980 a été réalisé dans une grande mesure à partir de registres. Ceux-ci ont fourni les données démographiques de base, celles sur le revenu et celles sur les études terminées (à l'exception des études faites à l'étranger). Pour compléter ces données,

dans le registre de population, les données sur les unités de logement contenues dans ce registre peuvent être reliées aux données qui se trouvent dans le registre de population et qui se rapportent aux occupants de ces unités. Autrement dit, les deux registres peuvent être séparés. On s'y prend de la même façon pour apparier le registre où sont inscrits le nom de l'employeur et le lieu de travail de chaque personne et le registre central des entreprises et des établissements et ainsi obtenir des données concernant la branche d'activité dans laquelle une personne travaille, ses déplacements pour aller travailler (navettage), etc.

3. LE RECENSEMENT DANS LES PAYS SCANDINAVES

Les quatre pays scandinaves possèdent des registres de population bien conçus du type décrit dans la section 2.1. Ils ont constitué, ou ont l'intention de le faire, des registres centraux des immeubles et des logements servant principalement à des fins administratives. Dans cette section, je vais décrire brièvement le genre de recensement effectué dans chaque pays et résumer l'orientation que cette activité est en train de prendre en Scandinavie.

3.1 Danemark

Le Danemark est le seul pays scandinave et, à ma connaissance, le seul pays d'Europe, qui a complètement renoncé à effectuer des recensements classiques pour n'en réaliser qu'à partir de registres. Ce changement s'est fait sur une période de plus de dix ans. Le registre central de population contenant des numéros de référence personnels a été créé en 1968 à des fins administratives et, en 1976, un recensement de la population (mais pas du logement) a été effectué à partir de ce registre. En 1977, on a créé un registre central des immeubles et des logements, toujours principalement dans un but administratif, et, en 1981, on a réalisé cette fois un recensement de la population et du logement, toujours à partir de registres. En 1979-1980, une mesure importante de plus a été prise, celle de demander aux employeurs d'ajouter un renseignement aux déclarations qu'ils envoient au fisc et dans lesquelles ils inscrivent les gains de chaque employé: ceux qui avaient plus d'un établissement devaient indiquer le lieu de travail de chaque employé. Ce renseignement supplémentaire a été exigé à des fins purement statistiques, et le bureau de la statistique a dû déployer des efforts considérables pour obtenir la collaboration des déclarants.

Les registres que tient Danmarks Statistik à des fins statistiques, au nombre de 37 environ, permettent de produire, une fois par an ou plus dans certains cas, des statistiques sur la population, l'emploi, le navettage, le revenu, le logement et la construction, et ce pour les municipalités et parfois des régions plus petites encore. Toutefois, pour des questions de coût, il n'est pas possible de faire le genre d'analyse qu'on fait d'après les données de recensement à une fréquence le moins comparables: pour avoir des analyses semblables à celles qui ont suivi le recensement de 1981, il faudra attendre 1991, et encore la risquent-elles d'être plus restreintes.

Le passage à un recensement réalisé à partir de registres a été facilité par la réorganisation du bureau central de la statistique du Danemark en 1966. Le Danmarks Statistik s'est vu accorder une certaine indépendance vis-à-vis du gouvernement central, ce qui a peut-être rassuré le public relativement à la question de la confidentialité. Cet organisme a maintenant le droit d'exiger, et d'utiliser à des fins statistiques, des données recueillies par les pouvoirs publics à des fins administratives et celui de participer à la création de registres contenant ce genre de données. Les problèmes auxquels Danmarks Statistik doit à présent faire face concernent principalement la qualité et l'actualité des données, qui sont toutes deux tributaires de l'efficacité de la procédure administrative. Ainsi, la lenteur avec laquelle les données des fichiers fiscaux ont été compilées (données sur la branche d'activité, la profession, le navettage et le revenu) a retardé l'analyse de ces données pour le recensement de 1981 jusqu'à l'été 1983. On s'attend par ailleurs que les statistiques sur la population active continueront d'avoir au moins un an de retard par rapport à l'année de référence à laquelle elles se rapportent. Il est particulièrement difficile

Les données administratives peuvent étayer un recensement *classique* de diverses façons (Redfern 1987, paragraphes 3.65-3.67), mais c'est leur utilisation dans un recensement *réalisé* à partir de *registres* qui sera le thème principal de cette communication. La section 2 décrit les registres dont on a besoin pour effectuer un tel recensement et la section 3 met en évidence les points communs et les différences entre les façons dont les quatre pays scandinaves ont choisi de procéder à cet égard. La section 4 examine ensuite les obstacles que les autres pays rencontreraient s'ils voulaient améliorer leurs systèmes de registres afin de pouvoir s'en servir pour réaliser des recensements, et il ressort de cette analyse que les problèmes soulevés sont davantage d'ordre administratif et politique que statistique.

Ce sont ces questions d'intérêt plus général qui constituent le second grand thème de la communication. La section 5 examine en détail la situation dans un pays où, pour des raisons politiques et idéologiques, les dossiers administratifs ne sont pas reliés entre eux au moyen d'un registre de population: le Royaume-Uni. La section 6 décrit une initiative récente en Australie, en vue d'améliorer les dossiers administratifs. Enfin, la section 7 résume les arguments politiques invoqués par les partisans et les adversaires du couplage des données administratives au moyen de registres de population et présente quelques raisons pour lesquelles les statisticiens devraient prendre une part prépondérante au débat sur la question.

2. REGISTRES NÉCESSAIRES À LA RÉALISATION D'UN RECENSEMENT

2.1 Registres de population

Pour réaliser un recensement à partir de registres, il faut, comme point de départ essentiel, un registre de population qui contienne des numéros de référence personnels et des adresses. Une correspondance biunivoque doit exister entre les numéros personnels et les membres de la population. Pour que le registre puisse être tenu à jour, les citoyens doivent également figurer dans les fichiers des divers organismes administratifs qui tiennent des dossiers afin de pouvoir servir à l'appariement de tous les dossiers à des fins statistiques.

La tenue d'un registre de population sert essentiellement à des fins administratives. Il s'agit d'une façon efficace d'organiser les nombreux rapports entre les pouvoirs publics, au niveau central et local, et le simple citoyen, pour ce qui a trait notamment aux impôts, à la sécurité sociale, aux services de santé offerts par l'état et à l'inscription sur les listes électorales. Pour être d'une utilité optimale, ces registres doivent servir à une vaste gamme d'activités administratives, de manière que les occasions de les mettre à jour et de les corriger soient fréquentes et que les citoyens s'habituent à donner leur numéro personnel.

La clé du système est le registre central de population dans lequel sont consignés des renseignements permettant d'identifier chaque personne (nom, date et lieu de naissance, date d'immigration, état matrimonial et éventuellement origine et citoyenneté) ainsi qu'un numéro de référence permanent. Dans la plupart des pays, le registre central de population contient les adresses les plus récentes, ce qui n'est toutefois pas le cas du registre français, le *Répertoire national d'identification des personnes physiques*. La fonction administrative première du registre central est de servir de point de référence pour les organismes administratifs qui peuvent y vérifier l'identité des personnes avec lesquelles ils ont des rapports et, s'il y a lieu, corriger ou inscrire les numéros de référence personnels dans leurs fichiers.

2.2 Autres registres importants

Pour réaliser un recensement de la population et du logement à partir de registres, on aussi recours à des registres fondés sur des unités autres que les personnes. Les plus importants sont les registres centraux des logements et les registres centraux des entreprises et des établissements (lieux de travail). Dans la mesure où le registre des logements attribue à chaque unité de logement (et pas seulement à l'immeuble ou à l'adresse) un code qui fait aussi partie de l'adresse inscrite

L'expérience européenne relative à l'utilisation des données administratives pour recenser la population: questions d'ordre politique

PHILIP REDFERN¹

RÉSUMÉ

L'expérience de quatre pays scandinaves fait ressortir les avantages et les inconvénients des recensements de la population réalisés à partir de registres et montre comment on pourrait remédier aux inconvénients. Dans d'autres pays, les tenants de cette façon de procéder se heurtent à des obstacles: soit on n'y possède pas les systèmes de données nécessaires ou de la qualité voulue, soit le public y voit une menace à la vie privée et s'interroge sur le pouvoir de l'État. Ces questions se situent bien au-delà du domaine de la statistique; elles sont d'ordre politique et administratif. Dans cette communication, la situation dans deux pays, le Royaume-Uni et l'Australie, est examinée. Au Royaume-Uni lorsque, par le passé, il a été question d'établir un registre de population en période de paix, ce genre de tentative a échoué, et l'opinion publique actuelle y est toujours hostile. Le gouvernement a néanmoins entrepris une réforme controversée des taxes locales, qui suppose la création de nouveaux registres. En Australie, le gouvernement déposait un projet de loi visant à introduire une carte d'identité nationale fondée sur un registre central et invoquait des arguments politiques clairs à l'appui de ce projet de loi; plus tard cette loi fut rétractée. La conclusion est que les questions soulevées par la réforme des systèmes de données méritent un examen approfondi, et quelques raisons pour lesquelles les statisticiens devraient prendre une part prépondérante au débat sont avancées.

MOTS CLÉS: Recensement de la population; cartes d'identité; numéro de référence personnel; registre de population; appariement de dossiers.

1. INTRODUCTION

Cette communication s'inspire d'une étude sur les différentes façons de procéder à un recensement de la population que j'ai réalisée pour le compte de l'Office statistique des communautés européennes (Redfern 1987). Pour cette étude, je me suis penché sur l'expérience des douze pays membres de la CEE et sur celle du Canada, de la Suède et des États-Unis. Il en est ressorti que les enquêtes par sondage peuvent compléter mais non remplacer les recensements pour la bonne raison qu'elles ne permettent pas d'obtenir des statistiques régionales fiables. L'utilisation d'un questionnaire long et d'un questionnaire abrégé au recensement du Canada et à celui des États-Unis est un exemple important de sondage complétant un dénombrement intégral. Il est question que la Norvège ait quant à elle recours à une enquête par sondage pour compléter les données tirées de registres se rapportant à la population entière (section 3.3). Il est possible de produire des données régionales à partir de registres contenant les adresses des personnes qui y sont inscrites. Si les domaines d'intérêt du recensement sont bien couverts par les registres (du point de vue des définitions, du champ d'observation, de l'exactitude et des périodes de référence) et que les registres peuvent être reliés entre eux, il est alors possible de créer, pour chaque individu, un enregistrement semblable à une déclaration de recensement et ainsi réaliser un recensement à partir des registres. Il s'agit essentiellement de recycler les données administratives à des fins statistiques. Les pressions que constituent les coûts et le fardeau de remplir les questionnaires habituels de recensement ont amené quatre pays scandinaves (le Danemark, la Finlande, la Norvège et la Suède) à adopter cette approche totalement ou en partie.

¹ Philip Redfern, 17 Fulwith Close, Harrogate, North Yorkshire, Angleterre, HG2 8HP.

- BROWNLEE, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*, (2^e éd.). New York: John Wiley and Sons.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley and Sons.
- DODGE, H.F., et ROMIG, H.G. (1959). *Sampling Inspection Tables*, (2^e éd.). New York: John Wiley and Sons.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. Vols. 1 et 2. New York: John Wiley and Sons.
- HENDERSON, R.H., et coll. (1973). Assessment of Vaccination Coverage, Vaccination Scar Rates, and Smallpox Scarring in Five Areas of West Africa. *Bulletin de l'Organisation mondiale de la Santé*, 48:183-194.
- HENDERSON, R.H., et SUNDARESAN, T. (1982). Cluster Sampling to Assess Immunization Coverage: A Review of Experience with a Simplified Sampling Method. *Bulletin de l'Organisation mondiale de la Santé*, 60:253-260.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- LEMESHOW, S., et coll. (1985). A Computer Simulation of the EPI Survey Strategy. *International Journal of Epidemiology*, 14, 3: 473-481.
- LEMESHOW, S., et ROBINSON, D. (1985). Enquêtes sur la couverture et l'impact des programmes: méthodes quantitatives utilisées par le Programme élargi de vaccination. *Rapport trimestriel de statistiques sanitaires mondiales*, 38, 1.
- LEMESHOW, S., HOSMER, D., et KLAR, J. (1987). *Sample Size Determination*. À être publié par l'Organisation mondiale de la Santé.
- LEVY, P.S., et LEMESHOW, S. (1980). *Sampling for Health Professionals*. Lifetime Learning Publications, New York: Van Nostrand Reinhold.
- ORGANISATION MONDIALE DE LA SANTÉ (1979). *Training for Mid-Level Managers. Evaluate Vaccination Coverage*. Genève, Programme élargi de vaccination de l'OMS, en collaboration avec le U.S. Department of Health and Human Services, Public Health Service, Center for Disease Control.
- SERFLING, R.E., et SHERMAN, I.L. (1975). Attribute Sampling Methods. Washington, D.C., U.S. Department of Health and Human Services, Public Health Service, Publication No. 1230.

BIBLIOGRAPHIE

La figure 3 montre la courbe caractéristique efficace pour ce plan d'échantillonnage partiel. Cette courbe permet de déterminer dans quelle probabilité un dispensaire sera classé correctement, étant donné un taux de couverture vaccinale. Nous supposons que les 294 dispensaires sont répartis uniformément entre les tranches de pourcentages et que le taux de couverture vaccinale des dispensaires d'un décile correspond à la valeur médiane pour ce décile. En nous servant de la CCE pour connaître la probabilité de classer un dispensaire parmi ceux ayant un taux de couverture vaccinale acceptable et en appliquant cette probabilité au nombre de dispensaires compris dans le décile correspondant, nous pouvons déterminer le nombre de dispensaires qui seront acceptés ou rejetés selon le critère établi (en l'occurrence, taux de couverture vaccinale supérieur à un pourcentage donné). Les résultats de cette projection sont présentés dans le tableau 3.

Un calcul rapide appliqué aux chiffres du tableau ci-dessus permet de constater que plus de 99% (183 sur 184) des dispensaires pour lesquels le taux de couverture vaccinale est inférieur à 70% seraient «rejetés» (c.-à-d. qu'ils seraient reconnus comme ayant un taux de couverture vaccinale insuffisant). Sur les 110 dispensaires pour lesquels le taux de couverture est supérieur à 70%, 62(56%) seraient reconnus, à juste titre, comme ayant un taux de couverture acceptable. Même si les 48 autres seraient classés par erreur parmi les dispensaires ayant un taux de couverture insuffisant, il convient de souligner que 63% d'entre eux (soit 30) ont un taux de couverture situé dans la tranche «marginale» (c.-à-d. 70 à 80%).

Les données relatives aux échantillons de dix enfants analysés au complet dans chacun des 294 territoires permettent, comme n'importe quel échantillon aléatoire stratifié, d'établir une estimation nationale. Suivant les mêmes hypothèses que celles utilisées pour le plan d'échantillonnage «classique», l'intervalle de confiance à 95% construit à l'aide de la méthode AQE permettrait d'estimer P (taux de couverture vaccinale à l'échelle nationale) à 1.8% près, soit un niveau de précision acceptable pour les besoins du directeur du PEV.

Il convient aussi de souligner que le nombre total d'enfants qui seraient examinés dans chaque territoire varierait de 10 à 24. De fait, compte tenu de la distribution des taux de couverture vaccinale supposée dans cet exemple, l'échantillon initial suffirait pour classer la majeure partie des dispensaires (autrement dit, environ 98% des 184 dispensaires ayant un taux de couverture inférieur à 70% seraient «rejetés» sur la seule base de l'échantillon initial $n_1:d_1 = 10:0$). Parmi le peu de dispensaires qui ne pourraient être classés sur la seule base de cet échantillon, quelques-uns seulement exigeraient l'examen des 14 enfants du second échantillon. Ainsi, la taille «moyenne» des échantillons prélevés dans les 294 territoires desservis par un dispensaire serait sensiblement inférieure à $n_1 + n_2$.

En conclusion, le CEQL peut être utile dans les situations où l'échantillonnage aléatoire stratifié classique (qui exige des échantillons suffisamment grands dans chaque strate pour que l'on ait des intervalles de confiance significatifs pour les estimations obtenues) est trop coûteux ou demande trop de temps. De fait, le CEQL n'est rien de plus qu'un autre moyen d'interpréter des données obtenues à l'aide d'un échantillon aléatoire stratifié sauf que là, les échantillons sont trop petits pour donner des intervalles de confiance significatifs. Comme le prélevement de petits échantillons peut se faire plus fréquemment, on pourrait penser à mettre sur pied un système par lequel on assurerait le suivi d'activités; cette tâche pourrait s'ajouter à celles accomplies sur le terrain, après que le personnel concerné aurait reçu une formation de base. Un échantillonnage plus fréquent pourrait avoir un autre avantage: au lieu de concentrer leur attention sur une cohorte qui est censée avoir déjà reçu tous les vaccins prévus, les directeurs de programme pourraient demander aux sondeurs de recueillir des données sur les enfants qui sont sur le point d'être vaccinés, c.-à-d. de déterminer si les enfants ont reçu les vaccins prévus pour leur âge. Cela permettrait de recueillir de l'information sur des activités plus récentes et de prendre les mesures nécessaires en vue d'accroître le taux de couverture vaccinale. Même si les intervalles de confiance fournissent beaucoup plus de renseignements pour qu'une simple décision dichotomique, il faudra probablement de très grands échantillons pour atteindre un niveau de précision intéressant dans le cas de strates relativement petites. Dans de telles circonstances, l'AQE est une solution qui mériterait d'être envisagée.

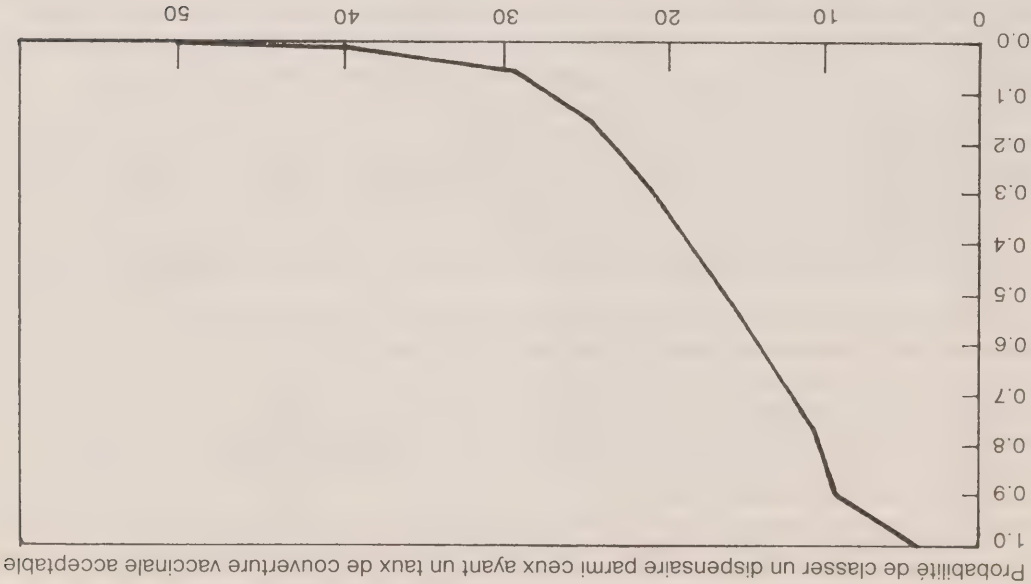


Figure 3: Courbe caractéristique efficace pour un plan d'échantillonnage double où $n_1:d_1 = 10:0$ et $n_2:d_2 = 14:3$

Pourcentage d'enfants n'ayant pas été vaccinés

D'après Dodge et Romig (1959) - Annexe 2: Courbes caractéristiques efficaces pour tous les plans d'échantillonnage double - ($N=51-100$)

Tableau 3

Répartition prévue de 294 dispensaires selon un plan d'échantillonnage double $n_1:d_1 = 10:0$ et $n_2:d_2 = 14:3$

Taux de couverture vaccinale sur le territoire du dispensaire (%)	Nombre de dispensaires	le taux de couverture vaccinale est supérieur à 70%	le taux de couverture vaccinale est égal ou inférieur à 70%
20- 30%	36	0	36
31- 40%	37	0	37
41- 50%	37	0	37
51- 60%	37	0	37
61- 70%	37	1	36
71- 80%	37	7	30
81- 90%	37	21	16
91-100%	36	34	2
Total	294	63	231

Nombre de dispensaires où le taux de couverture vaccinale est égal ou inférieur à 70% = 184.

Nombre de dispensaires classés correctement = 183 (99%).

Nombre de dispensaires où le taux de couverture vaccinale est supérieur à 70% = 110.

Nombre de dispensaires classés correctement = 62 (56%).

4. EXEMPLE D'APPLICATION DE L'ÂGE

Notre exemple illustre une situation qui ressemble à celle observée au Costa Rica; il porte sur les services de vaccination offerts aux enfants dans 294 dispensaires qui desservent la population du pays. Le directeur du PEV voudrait connaître le pourcentage d'enfants de 12 à 23 mois qui ont reçu tous les vaccins qui devaient leur être inoculés durant leur première année de vie. D'après les rapports fournis par le personnel des dispensaires, le directeur croit que le taux de couverture vaccinale pour l'ensemble du pays se situe autour de 60% mais les taux rapportés par chacun des 294 dispensaires varient de 20 à 100%; ces taux sont, croit-on, distribués uniformément. Le directeur du PEV soupçonne que les estimations indiquées dans les rapports peuvent ne pas être tout à fait exactes à cause d'erreurs touchant le numérateur et le dénominateur. Il décide par conséquent de sonder la population desservie par les dispensaires afin d'estimer le taux de couverture vaccinale dans chacun des 294 territoires de sorte qu'il soit possible ensuite de concentrer les efforts là où le taux de couverture vaccinale est relativement faible.

Pour les besoins de cette enquête, le directeur du PEV envisage tout d'abord un plan d'échantillonnage aléatoire stratifié «classique». On doit établir des estimations du taux de couverture vaccinale pour chacun des 294 dispensaires et les bornes de l'intervalle de confiance de ces estimations ne doivent pas excéder 10% en valeur absolue, étant donné $\alpha = 0.05$. Comme la population moyenne desservie par un dispensaire est d'environ 2500 et que 3.5% de cette population (selon des estimations) est constituée d'enfants de 12 à 23 mois, on estime à 88 (2500x0.035) le nombre d'enfants parmi lesquels un échantillon sera prélevé dans chaque territoire desservi par un dispensaire. La formule qui permet de calculer la taille de l'échantillon, et qui comprend un facteur de correction pour population finie, est définie dans Cochran (1977, p. 75); après application de cette formule, on obtient $n = 47$.

Ainsi, 47 (53%) des 88 enfants âgés de 12 à 23 mois feront partie de l'enquête dans chacun des 294 territoires. En tout, 13,818 enfants de ce groupe d'âge formeront l'échantillon global. En ce qui concerne le taux de couverture vaccinale estimé à l'échelle nationale, P peut être estimée à 0.5% près (en supposant le pire taux de couverture en ce qui a trait à la précision (50%)) et peu de variation entre les populations des divers territoires).

Le directeur considère ensuite une méthode AOE. Selon lui, tout dispensaire pour lequel on observe un taux de couverture vaccinale de 70% ou moins donne un rendement insatisfaisant et doit faire l'objet d'une supervision accrue. Le directeur veut être en mesure de reconnaître un dispensaire avec un taux de couverture de 70% dans une probabilité d'environ 0.95 et des dispensaires ayant des taux de couverture moins élevés dans une probabilité encore plus forte. Il considère plusieurs plans et propose finalement un échantillonnage double.

Le plan d'échantillonnage retenu peut être désigné comme suit: $n_1:d_1 = 10:0$ et $n_2:d_2 = 14:3$. Cela signifie que pour chaque territoire desservi par un dispensaire, on examinera un premier échantillon formé de dix enfants. Peu importe le nombre d'enfants que l'on trouvera non vaccinés dans cet échantillon, les dix seront examinés. Les données relatives au nombre d'enfants non vaccinés dans chaque échantillon serviront à établir des estimations pour des groupes de territoires et, finalement, à estimer le taux de couverture vaccinale à l'échelle du pays. Si on ne trouve aucun enfant non vacciné parmi les dix du premier échantillon, le dispensaire correspondant sera classé parmi les dispensaires ayant un taux de couverture vaccinale «acceptable». Si, dans ce même échantillon, on trouve au moins 4 enfants non vaccinés, le dispensaire correspondant sera classé parmi les dispensaires ayant un taux de couverture vaccinale «insuffisant». Dans l'un et l'autre cas, il n'est pas nécessaire de procéder à un autre échantillonnage. Toutefois, si l'examen du premier échantillon révèle 1, 2 ou 3 enfants non vaccinés, on prélève un second échantillon formé de 14 autres enfants. Dès que l'on a trouvé en tout 4 enfants non vaccinés (en comptant ceux qui ont été trouvés dans le premier échantillon), le sondage cesse et le dispensaire en question est classé parmi les dispensaires ayant un taux de couverture vaccinale «insuffisant». Cependant, si on n'a pas trouvé plus de trois enfants non vaccinés dans les deux échantillons combinés, le dispensaire correspondant est classé parmi les dispensaires ayant un taux de couverture vaccinale «acceptable».

Outre le fait d'«accepter» ou de «rejeter» un lot, nous pouvons considérer les échantillons aléatoires simples prélevés dans chaque dispensaire comme un échantillon stratifié et nous pouvons ainsi construire une estimation de population globale.

Bien que dans le CBQL, la valeur de n pour chaque strate soit trop faible pour produire de bons intervalles de confiance pour les estimations de strate, un plan d'échantillonnage bien conçu peut être un moyen de tester continuellement les strates et de les classer comme «acceptables» ou «inacceptables» par rapport à un critère particulier. Cela découle du fait que dans le CBQL les échantillons sont relativement petits et qu'il est donc plus probable que l'échantillonnage pourrait être plus fréquent. Un des avantages du CBQL est que les règles d'échantillonnage sont faciles à suivre, exigeant peu de connaissances additionnelles de la part du sondeur/classificateur. Enfin, comme les échantillons prélevés selon le CBQL sont, de fait, des échantillons aléatoires stratifiés, on peut grouper les résultats relatifs aux strates afin d'obtenir des estimations suffisamment précises pour des groupes de strates, par exemple des districts, des régions ou encore un pays.

Il est nécessaire de mettre en balance les avantages qui peuvent découler de l'utilisation du CÉQL et la perte de précision due aux petits échantillons prélevés dans chaque strate. La meilleure façon de juger de l'utilité du CÉQL est probablement d'étudier un exemple où l'échantillonnage aléatoire stratifié classique est comparé à l'échantillonnage fait selon le CÉQL.

Tableau 1
Valeurs de d* pour diverses combinaisons de P_o et n, alpha ≤ 0.01, 0.05, ou 0.10

n	$P_o, \alpha \leq 0.01$			$P_o, \alpha \leq 0.05$			$P_o, LPH \leq 0.10$			
	0.50	0.60	0.70	0.80	0.90	0.50	0.60	0.70	0.80	0.90
5	×	×	0	1	2	0	0	1	2	3
6	×	0	0	1	2	0	0	1	2	3
7	0	0	1	2	3	0	0	1	2	3
8	0	1	1	2	3	0	1	2	3	4
9	0	1	2	3	4	1	1	2	3	4
10	0	1	2	3	4	1	1	2	3	4
11	0	1	2	3	4	1	1	2	3	4
12	0	1	2	3	4	1	1	2	3	4
13	0	1	2	3	4	1	1	2	3	4
14	0	1	2	3	4	1	1	2	3	4
15	0	1	2	3	4	1	1	2	3	4
16	0	1	2	3	4	1	1	2	3	4
17	0	1	2	3	4	1	1	2	3	4
18	0	1	2	3	4	1	1	2	3	4
19	0	1	2	3	4	1	1	2	3	4
20	0	1	2	3	4	1	1	2	3	4

× Pas de test dans ce cas.

× Pas de test dans ce cas.

Tableau 2
Taille de l'échantillon et règle de décision pour le CEQL, Alpha = 0.05, Beta = 0.20, test unilatéral

P _a	0.50		0.60		0.70		0.80		0.90	
	n	d*	n	d*	n	d*	n	d*	n	d*

0.05	5	0	×	5	0	×	×	×	×	×
0.10	8	1	5	7	×	×	×	×	×	×
0.15	11	2	9	1	×	×	×	×	×	×
0.20	15	3	12	2	×	×	×	×	×	×
0.25	23	7	16	3	×	×	×	×	×	×
0.30	37	13	24	5	×	×	×	×	×	×
0.35	67	26	38	10	×	×	×	×	×	×
0.40	153	66	86	17	×	×	×	×	×	×
0.45	617	288	340	33	×	×	×	×	×	×
0.50	151	80	67	80	×	×	×	×	×	×
0.55	601	340	137	86	×	×	×	×	×	×
0.60	137	86	62	37	×	×	×	×	×	×
0.65	535	356	29	19	×	×	×	×	×	×
0.70	109	80	19	11	×	×	×	×	×	×
0.75	419	321	10	5	×	×	×	×	×	×
0.80	69	58	6	3	×	×	×	×	×	×
0.85	253	219	10	6	×	×	×	×	×	×

× Taille de l'échantillon inférieure à 5.

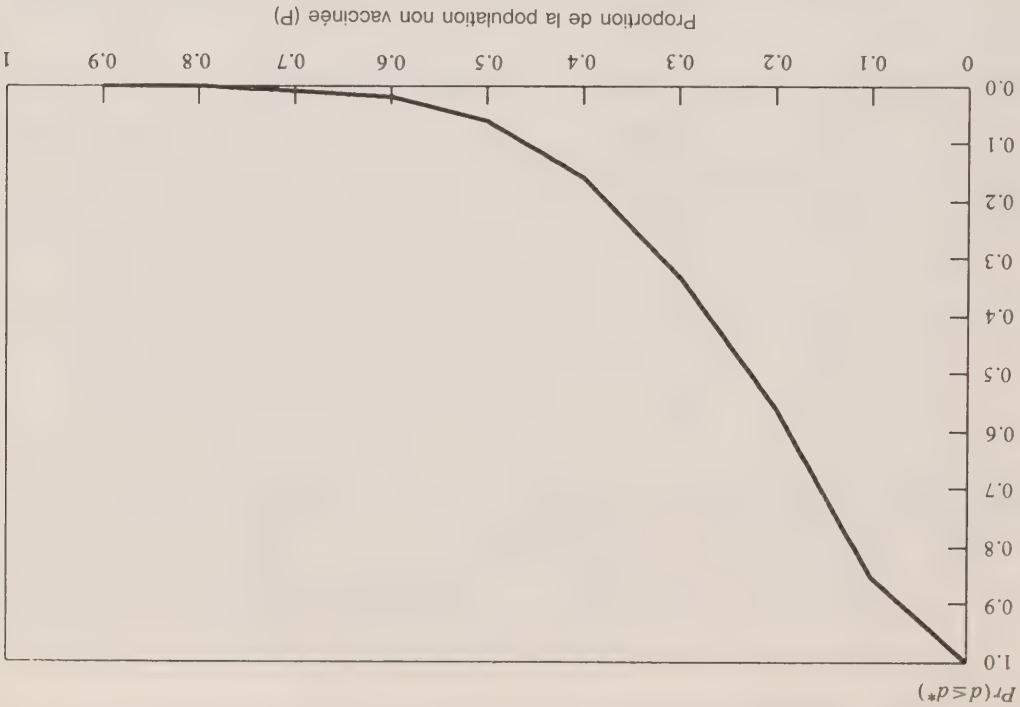


Figure 2: Courbe caractéristique efficace pour $n=7$ et $d^*=1$

que $Pr(d \leq d^*) = \alpha$, nous devons calculer $Pr(d \leq d^*)$ pour un certain nombre de valeurs de d^* . De toute évidence, si nous choisissons $d^* = 1$, nous obtenons $Pr(d \leq d^*) = 0.0625$ et la puissance du test (si 70% de la population n'est pas vaccinée) est égale à 0.0038.

Les résultats d'une telle analyse peuvent être représentés par une **courbe caractéristique efficace (CCE)**, qui met en relation la proportion P de la population qui n'est pas vaccinée (axe horizontal) et la probabilité de rejeter l'hypothèse nulle $H_0: P = P_0$ (axe vertical) et de conclure que le taux de couverture vaccinale est acceptable. Chaque paire de valeurs (n, d^*) produira sa courbe propre. La figure 2 illustre une CCE typique pour $n = 7$ et $d^* = 1$.

Le sondeur choisira habituellement une valeur de d^* pour laquelle l'erreur de première espèce sera inférieure à α . Cette précaution se traduit parfois par un test extrêmement prudent. Par exemple, si $n = 7$, $d^* = 0$ et $P_0 = 0.5$, α sera égale à 0.0078. Dans le cas qui nous occupe, il serait plus juste d'opter pour une valeur $d^* = 1$ (comme dans la figure 2), ce qui donne une valeur $\alpha = 0.0625$. Le tableau 1 donne les valeurs de d^* pour de petits échantillons ($n \leq 20$) de telle manière que α n'excède pas la probabilité d'erreur de première espèce indiquée (0.01, 0.05 ou 0.10) pour diverses combinaisons de n et de P_0 . Pour des renseignements détaillés sur la construction de ce tableau, voir Dodge et Romig (1959).

Le choix du plan d'échantillonnage se réduit essentiellement à combiner la puissance $1 - \beta$ voulue avec le niveau α voulu. Au lieu de présenter des courbes que l'on peut difficilement lire avec précision, nous avons construit un tableau de valeurs de paires (n, d^*) pour $\alpha = 0.05$, $\beta = 0.20$ et certaines valeurs de P suivant l'hypothèse nulle (P_0) et l'hypothèse alternative (P_a) (voir tableau 2). Dans ce tableau, les paires (n, d^*) ont été déterminées de telle manière que $Pr(d \leq d^* | n, P_0) \leq \alpha$ et $Pr(d \leq d^* + 1 | n, P_0) > \alpha$. (1987).

Ce tableau illustre bien le compromis qu'il est nécessaire de faire entre la puissance et la taille d'échantillon dans le CÉQL. Par exemple, il est essentiellement impossible d'utiliser un échantillon de taille $n = 5$ avec des valeurs $\alpha = 0.05$ et $\beta = 0.20$ à moins que P_0 soit effectivement près de 0. Par conséquent, les sondeurs qui disposent de ressources limitées doivent être prêts à transiger sur la valeur de β ou la différence entre P_0 et P_a .

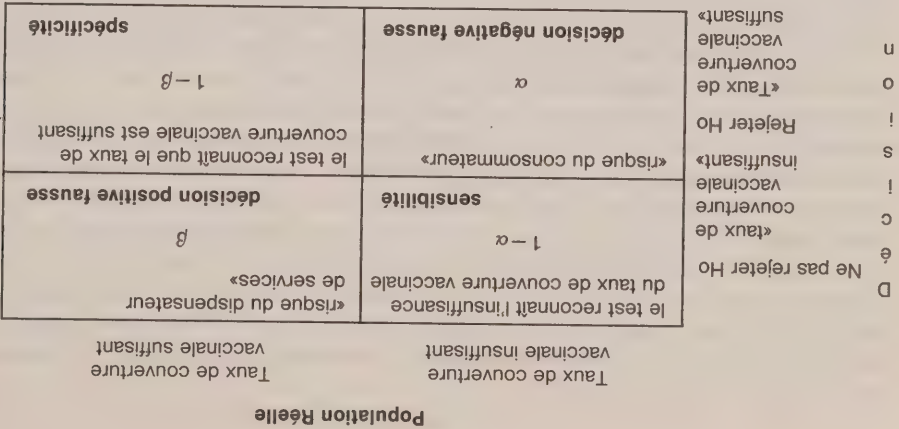


Figure 1. Conséquences du test d'hypothèses dans le CÉQL (contrôle par échantillonnage de la qualité des lots)

prédisposés. Par conséquent, le fait de déclarer que la population est immunisée dans une proportion acceptable alors que ce n'est pas le cas risque d'avoir des conséquences graves. Par ailleurs, l'erreur de seconde espèce (rejet d'un lot acceptable) n'est pas aussi grave puisqu'une décision positive fausse amènera le personnel de santé à concentrer son attention sur une population qui est déjà suffisamment protégée.

La difficulté première avec le CÉQL n'est pas tant de déterminer la taille de l'échantillon que de trouver un juste équilibre entre la taille de l'échantillon et la région de rejet. Dans tous les cas, le calcul de β dépendra de la valeur réelle de P lorsque celle-ci est supposée être différente de P_0 .

Dans la pratique, on fixera au départ un niveau minimum de prestation de services en se fondant sur la répartition probable des niveaux de prestation entre les lots et sur le degré de faiblesse (c.-à-d. définir un niveau qui soit réaliste). Une fois ce niveau déterminé, on considérera diverses tailles d'échantillons par rapport au nombre de lots qui pourraient être mal classés par suite d'une erreur de première ou de seconde espèce. Si la taille d'échantillon retenue est trop élevée pour être utile, plusieurs solutions peuvent être envisagées: a) conserver le plan d'échantillonnage mais réduire la fréquence des échantillonnages; b) choisir un autre niveau critique, qui permettra d'utiliser un échantillon de taille moindre; c) choisir un autre plan d'échantillonnage du type AQE (comme l'échantillonnage double ou l'échantillonnage progressif), qui viserait à classer les lots tout en conservant sa fonctionnalité; d) mettre de côté l'AQE. Le calcul des probabilités et la détermination des tailles d'échantillon nécessaires peuvent se faire à l'aide de la distribution binomiale. Nous supposons comme d'habitude que N est très grand par rapport à n , avec une valeur de N élevée, on peut substituer en pratique la distribution de Poisson à la distribution binomiale. Cependant, si jamais N n'est pas très élevé par rapport à n , on peut se servir de la distribution hypergéométrique, comme l'indique Brownlee (1965, sec. 3.15). Si nous définissons p comme la probabilité d'observer la caractéristique, alors la probabilité de trouver exactement d personnes possédant la caractéristique dans un échantillon de taille n est donnée par l'équation

$$p(d) = \binom{n}{d} p^d (1-p)^{n-d}.$$

Supposons que nous décidions d'utiliser un échantillon de taille $n = 7$. La région de rejet pour le test précise que nous devrions rejeter H_0 (et reconnaître que le «lot» est immunisé dans une proportion satisfaisante) si $d \leq d^*$. Pour savoir quelle doit être la valeur de d^* de telle manière

2. LA MÉTHODE AQE

La méthode AQE trouve son origine dans l'échantillonnage et l'inspection des produits manufacturés (Dodge et Romig 1959); on avait alors mis au point cette méthode dans le but de maintenir le plus bas possible les coûts de main-d'oeuvre et les autres coûts liés à l'échantillonnage. Une variante de l'AQE, soit le contrôle par échantillonnage de la qualité des lots (CEQL), ressemble à l'échantillonnage stratifié mais les échantillons obtenus sont trop petits pour produire des intervalles de confiance que l'on peut normalement qualifier de suffisamment étroits pour des estimations relatives à une strate particulière (appelée habituellement «lot» dans l'industrie). Au lieu de cela, on juge de la qualité d'un lot en fonction de la probabilité que le nombre d'articles défectueux dans l'échantillon soit égal ou inférieur à un nombre donné. On peut combiner les résultats des échantillons tirés de tous les lots entiers s'excluant mutuellement afin d'obtenir une estimation globale précise de la qualité moyenne du produit.

Qu'il soit question de fabrication de produits ou de soins de santé, les objectifs et les modalités d'application de l'AQE sont les mêmes. L'acheteur de biens n'acceptera pas un lot dans lequel le nombre d'articles défectueux dépasse un certain pourcentage (P_1) tandis que le fabricant surveillera constamment la production dans le but de repérer les produits pour lesquels le pourcentage d'articles défectueux sera supérieur au pourcentage prévu (P_2). Il n'est pas rare que P_1 et P_2 soient différents. On peut voir facilement la similitude qui existe entre les objectifs d'un fabricant et ceux d'un superviseur de programme de soins de santé. Dans le premier cas, on produit des articles alors que dans le second cas, on «produit» des enfants vaccinés.

De façon générale, un lot est une unité «utile au point de vue des opérations». Par exemple, si, dans une application industrielle, plusieurs machines produisent la même pièce et que trois opérateurs sont affectés à chaque machine, il est possible de choisir des «lots» provenant de la même machine — surtout si les erreurs de fabrication sont plus souvent attribuables à une défaillance mécanique qu'à l'opérateur.

En ce qui concerne les services de santé publique, un gestionnaire pourrait définir un «lot» comme un groupe de personnes qui reçoivent des services d'une unité opérationnelle — par exemple l'équipe de vaccination d'un dispensaire — durant une période déterminée. On peut faire coïncider les séances d'échantillonnage avec les périodes où les maladies contre lesquelles il existe un vaccin sont plus susceptibles de faire des ravages; cependant, la fréquence des échantillonnages dépendra plus souvent qu'autrement des délais et des coûts liés à cette opération.

Dans le domaine de la santé publique, on commet une erreur grave lorsqu'on juge qu'une population est suffisamment protégée («accepter le lot») alors qu'en réalité, elle ne l'est pas. Afin de tenir compte de cette éventualité, nous allons imaginer un test unilatéral.

L'hypothèse nulle (avec critère de 50%) est

$$H_0: P \geq P_0 \text{ (c.-à-d., proportion d'enfants non vaccinés } \geq 0.50)$$

et l'hypothèse alternative

$$H_a: P < P_0 \text{ (c.-à-d., proportion d'enfants non vaccinés } < 0.50).$$

Le tableau à quatre cases de la figure 1 expose les conséquences du test. Comme il s'agit d'un test unilatéral et que nous supposons que la population n'est pas suffisamment protégée à moins que nous rejétions H_0 , l'erreur de première espèce (accepter le lot alors qu'il renferme un trop grand nombre d'articles défectueux — décision négative fausse) est la plus grave que nous puissions commettre. Par exemple, si nous tenons pour acquis qu'une population (lot) d'enfants est vaccinée dans une proportion acceptable alors qu'en réalité, il n'en est rien, les risques de contagion sont plus élevés du fait que la population compte un plus grand nombre d'enfants

par 7», concernent ordinairement 30 grappes et 7 personnes par grappe. De fait, l'intérêt indéniable de la méthode d'enquête utilisée dans le PEV tient à la simplicité du plan de sondage, aux règles d'application standardisées et à la façon simple de dépouiller et d'interpréter les résultats. Une analyse et une critique de la méthode sur le plan théorique existent dans d'autres ouvrages (Lemeshow et coll. 1985 et Robinson 1985).

Récemment, des agents du PEV ont reconnu plusieurs lacunes à la méthode d'enquête. La première est que les résultats obtenus par cette méthode sont relativement imprécis — il peut y avoir jusqu'à 10 points de pourcentage entre le taux de couverture vaccinale estimé et le taux de couverture réel dans la population sondée. Dans les pays en voie de développement qui ont pu atteindre un taux de couverture vaccinale élevé, la méthode est trop imprécise pour faire ressortir les variations significatives qui surviennent entre deux enquêtes ou les différences entre les strates d'une population étudiée.

La seconde lacune des enquêtes réalisées dans le cadre du PEV est que même si elles sont relativement faciles à réaliser, elles représentent une opération encore trop gigantesque pour que les gestionnaires locaux puissent y recourir afin d'évaluer les opérations dans leur domaine de compétence. C'est pourquoi ce genre d'enquêtes portent, encore aujourd'hui, sur la population entière d'un pays ou des groupes de population relativement grands (par exemple, des groupes dont l'effectif se chiffre en millions). Bien que les résultats de ces enquêtes soient utiles pour les gestionnaires des niveaux supérieurs, les gestionnaires locaux et les superviseurs ne sont pas en mesure de s'en servir à leur niveau.

Les enquêtes du PEV servent habituellement à évaluer le pourcentage d'enfants d'une cohorte (enfants âgés habituellement de 12 à 23 mois) qui seraient censés avoir reçu la série complète de vaccins prévus par le PEV. La troisième lacune des enquêtes réalisées en vertu de ce programme est qu'elles servent à mesurer des opérations qui ont eu lieu il y a plus d'un an; il peut y avoir eu des changements considérables dans l'intervalle.

Enfin, un autre objectif du PEV est de mettre sur pied un système d'enregistrement fiable qui peut servir à évaluer le taux de couverture vaccinale et à en suivre l'évolution, les enquêtes étant le principal moyen de vérifier la justesse des enregistrements. Toutefois, compte tenu des groupes d'âge visés par l'enquête, il est souvent difficile de trouver la série d'enregistrements qui coïncident avec la période où les enfants concernés ont été vaccinés.

Dans cet article, nous proposons une méthode qui permet de suivre l'évolution d'un programme de soins de santé et d'établir si les opérations se situent à un niveau acceptable donné. À cette fin, nous utilisons un type particulier d'échantillonnage aléatoire stratifié (Cochran 1977, Hansen et coll. 1953; Kish 1965; Levy et Lemeshow 1980) qui produit de très petits échantillons à partir d'unités de la population définies au point de vue opérationnel. Non seulement ce genre d'échantillonnage axé sur la collectivité permettra de contrôler les opérations de programmes au sein de populations relativement petites ou dans de petits territoires, mais encore il permettra aux gestionnaires de tous les niveaux hiérarchiques de presque d'obtenir des estimations qui serviront à évaluer continuellement les opérations de programmes avec suffisamment de précision. Dans les régions où ont été élaborés des systèmes d'enregistrement pouvant servir à contrôler les opérations de programmes, on peut utiliser la même méthode d'échantillonnage pour confirmer les enregistrements et vérifier si ceux-ci produisent un énumérateur et un dénominateur exacts. Une fois confirmés, ces enregistrements peuvent être considérés comme la principale source d'information aux fins du suivi et de l'évaluation des programmes. Cette méthode d'échantillonnage, que nous proposons de substituer aux méthodes plus traditionnelles utilisées dans l'évaluation des programmes de santé publique, est désignée par l'expression générale suivante : assurance de la qualité par échantillonnage (AQE) — expression bien connue dans les domaines du génie, de la fabrication et du commerce.

Assurance de la qualité par échantillonnage pour l'évaluation des paramètres de santé dans les pays en voie de développement

STANLEY LEMESHOW et GEORGE STROH, JR.¹

RÉSUMÉ

Une des principales tâches des agents sanitaires en poste dans les pays en voie de développement est de vérifier si une population répond à certaines normes, par exemple la proportion d'habitants vaccinés contre une certaine maladie. Comme les populations sont généralement nombreuses et que les ressources et le temps nécessaires à des études sont limités, on doit habituellement procéder par échantillonnage et établir des estimations pour la population toute entière. Selon la proportion de personnes non vaccinées dans l'échantillon, on déterminera si la couverture vaccinale est acceptable ou s'il y a lieu d'accroître les efforts pour accroître cette couverture. Plusieurs méthodes d'échantillonnage sont actuellement en usage. Parmi celles-ci figure une version modifiée de la méthode d'échantillonnage en grappes, recommandée par le Programme élargi de vaccination (PEV) de l'Organisation mondiale de la Santé. Plus récemment, on a suggéré que l'assurance de la qualité par échantillonnage (AQE), une méthode couramment utilisée dans l'inspection des produits manufacturés, pourrait être utile au suivi des programmes de santé. Dans cet article, nous décrivons la méthode AQE et donnons un exemple d'application.

MOTS CLÉS: Échantillonnage des lots; assurance de la qualité; échantillonnage d'acceptation; couverture vaccinale.

1. INTRODUCTION

Une des préoccupations constantes des directeurs de programmes de soins de santé est d'élaborer et d'appliquer des méthodes pratiques et efficaces pour contrôler et évaluer les opérations. Dans les pays en voie de développement, cette tâche est habituellement complexe car les dossiers sont rarement à jour, les rapports des établissements de santé répartis ici et là sont ordinairement remis en retard ou ne sont pas remis du tout et la taille exacte des populations cibles n'est pas connue. En conséquence, les enquêtes menées auprès des collectivités sont souvent le seul moyen d'obtenir des données fiables concernant le nombre de personnes qui présentent une caractéristique (numérateur) et le nombre de personnes étudiées (dénominateur). Cependant, ce genre d'enquêtes peuvent être difficiles à organiser et à réaliser et sont souvent trop coûteuses pour servir au suivi des opérations de programmes.

Le Programme élargi de vaccination (PEV) de l'Organisation mondiale de la Santé (OMS) (voir Henderson et Sundaresan 1982) est probablement le meilleur exemple de programme où les enquêtes menées auprès de collectivités servent régulièrement à recueillir des données. Depuis sa création, le PEV utilise une méthode d'échantillonnage en grappes pour évaluer la couverture de vaccination chez les jeunes enfants (voir Serfling et Sherman 1975 et Henderson et coll. 1973). On a imaginé une méthodologie aussi simple que possible sur le plan théorique et sur le plan pratique de manière à permettre aux directeurs de programmes et aux superviseurs, qui souvent ont peu de connaissances dans les méthodes de sondage, d'organiser et de réaliser les enquêtes (voir OMS 1979). Ces enquêtes, que l'on appelle «30

¹ Stanley Lemeshow, Ph. D., est professeur de biostatistique et titulaire de la chaire de biostatistique et d'épidémiologie à la Division de santé publique de l'Université du Massachusetts, Amherst, MA. George Stroh Jr., M.P.H., Centers for Disease Control, Atlanta, GA.

BIBLIOGRAPHIE

- GROOTAERT, C. (1986). The use of multiple diaries in a household expenditure survey in Hong Kong. *Journal of the American Statistical Association*, 81, 938-944.
- HIRSHLEIFER, J. (1984). *Price Theory and Applications* (3e éd.). Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- KEMSELEY, W.F.F. (1961). The Household Expenditure Enquiry of the Ministry of Labour: Variability in the 1953-54 Enquiry. *Applied Statistics*, 10, 117-135.
- KEMSELEY, W.F.F., et NICHOLSON, J.L. (1960). Some experiments in methods of conducting family expenditure surveys. *Journal of the Royal Statistical Society, Series A*, 123, 307-328.
- LEWIS, H. F. (1948). A comparison of consumer responses to weekly and monthly purchase panels. *Journal of Marketing*, 12, 449-454.
- McKENZIE, J. (1983). The accuracy of telephone call data collected by diary methods. *Journal of Marketing Research*, 20, 417-427.
- NETER, J. (1970). Measurement errors in reports of consumer expenditures. *Journal of Marketing Research*, 7, 11-25.
- PARFITT, J. (1967). A comparison of purchase recall with diary panel records. *Journal of Advertising Research*, 7, 16-31.
- PEARL, R.B. (1968). Methodology of Consumer Expenditure Surveys. Technical Working Paper 27, Washington D.C.: U.S. Bureau of the Census.
- SANDAGE, C.H. (1956). Do research panels wear out? *Journal of Marketing*, 20, 397-401.
- SODOL, M.G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 59, 52-68.
- STANTON, J.L., et TUCCI, L.A. (1982). The measurement of consumption: A comparison of surveys and diaries. *Journal of Marketing Research*, 19, 274-277.
- SUDMAN, S. (1964a). On the accuracy of recording of consumer panels: I. *Journal of Marketing Research*, 1, 14-20.
- SUDMAN, S. (1964b). On the accuracy of recording of consumer panels: II. *Journal of Marketing Research*, 1, 69-83.
- SUDMAN, S., et FERBER, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research*, 11, 128-135.
- WIND, Y., et LERNER, D. (1979). On the measurement of purchase data: Surveys versus purchase diaries. *Journal of Marketing Research*, 16, 39-47.

5.6 Niveau d'instruction

Le lien de l'écart entre les valeurs de l'enquête préliminaire et celles du carnet avec le niveau d'instruction des répondants est illustré dans le tableau 6. Comme on peut le constater, la tendance à surestimer l'utilisation dans l'enquête préliminaire est une caractéristique commune aux répondants de tous les niveaux d'instruction. Les répondants ayant le niveau d'instruction le moins élevé tendent cependant à surestimer l'utilisation de l'interurbain à un degré moindre que les répondants plus instruits. On observe, dans les enquêtes, la tendance la plus prononcée à surestimer l'utilisation chez les répondants qui ont fait des études secondaires; viennent ensuite les répondants qui ont fait des études universitaires. Les résultats de l'analyse de la variance à un critère de classification et des tests *t* studentisés de Tukey indiquent cependant que les écarts pour les niveaux d'instruction ne sont pas statistiquement significatifs au niveau de $p = 0.05$.

5.7 Résultats de l'analyse de la variance à quatre critères de classification

Les effets principaux et les interactions des variables démographiques qui expliquent l'écart entre les valeurs de l'enquête préliminaires et celles du carnet sont présentées au tableau 7. On observe que le revenu et le sexe des répondants, et leur interaction, sont les variables ayant des valeurs *p* statistiquement significatives. Les autres effets principaux et interactions ne contribuent pas à expliquer les variations de l'écart.

6. CONCLUSION

Les résultats de notre étude indiquent qu'au niveau des répondants, les valeurs de l'enquête préliminaire sont une mesure très inexacte de l'utilisation réelle de l'interurbain. Nos résultats confirment les conclusions de Parfitt (1967), de Sudman (1964) et de Wind et Lerner (1982) qui ont étudié cette question pour des produits de consommation autres que des services. Nous ne pouvons toutefois affirmer que nos résultats confirment ou infirment ceux de Stanton et Tucci (1984) étant donné que les périodes de référence, et par conséquent les périodes pour lesquelles on fait appel à la mémoire des répondants, sont très différentes dans les deux études.

Nos résultats font plus que confirmer les conclusions d'études antérieures et à étendre l'analyse de produits à celle d'un service aux consommateurs. Ils mettent en lumière le fait que la surestimation que l'on observe lors des enquêtes-mémoire varie selon deux facteurs démographiques importants: le revenu du ménage et le sexe du répondant. Les répondants qui déclarent un très faible revenu tendent à sous-estimer l'utilisation de l'interurbain alors que l'on observe exactement le contraire chez les répondants dont les revenus sont plus élevés. En outre, ce rapport a tendance à augmenter de façon monotone à mesure que les revenus augmentent, puis à diminuer dans les tranches de revenu supérieures. Les femmes surestiment leur utilisation de l'interurbain dans une proportion beaucoup plus grande que les hommes. Considérés dans leur ensemble, les résultats de notre étude donnent à penser que la méthode de l'enquête-mémoire pour la collecte de renseignements sur les dépenses de consommation entraîne une forte possibilité de problèmes de mesure.

REMERCIEMENTS

Nous sommes reconnaissants aux arbitres anonymes de leurs précieuses suggestions qui ont permis d'améliorer la clarté de notre travail.

Tableau 6

Résultats de l'analyse de la variance à un critère de classification illustrant l'écart entre les valeurs de l'enquête préliminaire et celles du carnet en fonction du niveau d'instruction du répondant

Écarts selon le niveau d'instruction (enquête préliminaire — carnet)

Mois	Écarts selon le niveau d'instruction (enquête préliminaire — carnet)			
	Quelques années d'études secondaires	Études secondaires	Quelques années d'études universitaires	4 années d'études universitaires terminées
Février	0.790	1.015	0.951	0.578
Mars	0.290	0.853	0.592	0.059
Avril	0.556	1.275	0.979	0.445
Mai	0.685	1.134	0.756	0.345
Juin	0.548	1.158	1.111	0.696
Juillet	0.194	0.931	1.021	0.467
Août	0.347	0.891	0.845	0.620
Septembre	1.040	1.137	1.190	1.018
Octobre	0.468	1.195	0.826	0.878
Novembre	0.508	1.119	0.896	0.592
Décembre	0.081	0.500	0.244	0.061
Janvier	-0.097	0.626	0.842	0.129
Moyenne ^a	0.438	0.986	0.854	0.491
<i>n</i>	124	476	431	490

^a Moyenne des douze mois.
* Significative à un niveau de 0.01.
** Significative à un niveau de 0.05.

Tableau 7

Résultats de l'analyse de la variance à quatre critères de classification illustrant l'écart entre les valeurs de l'enquête préliminaire et celles du carnet sur l'utilisation de l'intervurbain en fonction de facteurs démographiques (sexe, niveau d'instruction, âge et revenu)

Variable	D.I.	Somme des carrés	Valeur F	Valeur <i>p</i>
Sexe	1	131.082	6.48	0.011**
Niveau d'instruction	3	79.001	1.30	0.272
Âge	3	58.465	0.96	0.409
Revenu	4	210.077	2.60	0.035**
Sexe et niveau d'instruction	3	77.629	1.28	0.280
Sexe et revenu	4	220.032	2.72	0.028**
Sexe et âge	3	47.311	0.78	0.506
Niveau d'instruction et revenu	12	263.931	1.09	0.367
Niveau d'instruction et âge	9	81.083	0.45	0.911
Revenu et âge	12	211.718	0.87	0.576

* Significative à un niveau de 0.01.
** Significative à un niveau de 0.05.

Tableau 5

Résultats de l'analyse de la variance à un critère de classification illustrant l'écart entre les valeurs de l'enquête préliminaires et celles du carnet sur l'utilisation de l'intervrain en fonction de l'âge du répondant					
Écarts selon l'âge (Enquête préliminaire — carnet)					
Mois	Moins de 31 ans	31-40	41-50	Plus de 50 ans	Valeur p
Février	0.632	0.749	1.026	0.949	0.310
Mars	0.016	0.348	0.837	0.709	0.210
Avril	0.413	1.083	1.174	0.889	0.209
Mai	0.305	0.706	1.085	0.923	0.217
Jun	0.525	0.845	1.570	0.989	0.080
Juillet	0.535	0.706	1.226	0.667	0.371
Août	0.507	0.807	1.070	0.667	0.578
Septembre	0.924	1.003	1.459	1.109	0.580
Octobre	0.789	0.816	1.307	0.903	0.583
Novembre	0.632	0.805	1.415	0.741	0.240
Décembre	0.337	0.203	0.574	— 0.030	0.494
Janvier	0.603	0.519	0.922	0.069	0.197
Moyenne ^a	0.518	0.716	1.139	0.715	0.385
n	383	374	270	495	

^a Moyenne des douze mois.
* Significative à un niveau de 0.01.
** Significative à un niveau de 0.05.

valeurs de l'enquête préliminaire sur l'estimation de l'utilisation peuvent, en raison des cir-constances présentes au moment de la consommation, différer des valeurs consignées dans le carnet et correspondant à l'utilisation réelle.

Lorsque les revenus des ménages augmentent, l'intervrain continue d'être considéré comme un service de luxe. Toutefois, alors que les répondants des tranches de revenu inférieures con-sidèrent les appels interurbains sont des dépenses non nécessaires (et peut-être même extravagantes), les répondants mieux nantis «prévoient» utiliser l'intervrain plus souvent que les autres moyens de communication. Ainsi, lorsqu'on leur demande de déclarer leur utilisati-on «prévue», ces répondants ont tendance à surestimer le nombre de leurs appels interur-bains puisque, dans la plupart des cas, il s'agit de leur moyen de communication préféré.

5.5 Âge

Le tableau 5 montre que les répondants de chaque groupe d'âge tendent à surestimer leur utilisation prévue par rapport à l'utilisation «réelle» consignée dans leur carnet. Bien que les valeurs p de l'analyse de la variance à un critère de classification n'indiquent la présence d'aucun lien significatif entre les méthodes de collecte des données et l'âge des répondants, pour 10 des 12 mois les répondants de moins de 31 ans avaient un plus faible écart que les répondants plus âgés. En outre, les écarts moyens pour les répondants de 31 à 40 ans et pour ceux de 50 ans et plus étaient plus faibles que la moyenne pour les moins de 31 ans pour chacune des douze périodes. Par conséquent, le lien entre les écarts dans les valeurs de l'enquête préliminaire et celles du carnet et l'âge du répondant est une fonction croissante monotone jusqu'à 50 ans, à partir duquel l'écart, bien que toujours positif, commence à diminuer. Encore une fois, les écarts moyens dans les divers groupes d'âge ne sont pas statistiquement significatifs selon l'analyse de la variance à un critère de classification et les tests t studentisés de Tukey.

Résultats de l'analyse de la variance illustrant l'écart entre les valeurs de l'enquête préliminaire et celles du carnet sur l'utilisation de l'interrubain en fonction du revenu du répondant

Tableau 4

Écarts selon le revenu (Enquête préliminaire — carnet)					
	0-\$5,000	\$5,001 – 10,000	\$10,001 – 15,000	\$15,001 – 20,000	Plus de 20,000
	(1)	(2)	(3)	(4)	(5)
Février	–0.010	–0.583	1.180	1.120	0.571
Mars	–0.480	–0.738	0.780	1.009	–0.062
Avril	–0.337	0.851	1.188	1.258	0.550
Mai	–0.327	0.560	0.928	0.991	0.636
Juin	0.102	0.911	1.027	1.331	0.756
Juillet	–0.439	0.500	0.895	1.050	0.694
Août	–0.408	0.512	1.021	1.235	0.498
Septembre	0.306	0.798	1.298	1.367	0.976
Octobre	–0.469	0.542	1.231	1.413	0.720
Novembre	0.010	0.494	0.941	1.214	0.741
Décembre	–1.010	0.060	0.209	0.792	0.101
Janvier	–0.633	–0.339	0.654	0.956	0.392
Moyenne ^{a,b}	–0.308	0.517	0.946	1.145	0.548
n	98	168	373	341	536

a Moyenne des douze mois.
b Test de contraste de Tukey: (1) et (4) et (1) et (3) sont différents au niveau de $p = 0.05$.
* Significative à un niveau de 0.01.
** Significative à un niveau de 0.05.

5.4 Revenu

Dans le tableau 4, nous montrons l'écart entre les valeurs de l'enquête préliminaire et celles du carnet en fonction du revenu du ménage des répondants. Pour 6 des 12 mois, les valeurs p de l'analyse de la variance à un critère de classification sont statistiquement significatives au niveau de 0.05 ou moins; elles le sont au niveau de 0.037 pour la moyenne des 12 mois. De plus, les résultats du test t studentisé de Tukey indiquent que les répondants dont le revenu annuel du ménage est dans la tranche 1 (\$5,000 ou moins) sont statistiquement distincts des répondants qui ont des revenus de l'ordre de \$10,000 à \$20,000.

On peut observer une particularité évidente dans les résultats présentés dans le tableau 4. Pour les répondants qui se trouvent dans la tranche de revenu la moins élevée (\$5,000 ou moins), l'estimation de l'utilisation mensuelle moyenne est inférieure à l'utilisation mensuelle réelle pour 9 des 12 mois. En outre, on constate que plus le revenu du ménage augmente plus on a tendance à surestimer l'utilisation, mais que cette tendance s'arrête dans la tranche de revenu la plus élevée. Il est possible que les consommateurs des tranches de revenu inférieures considèrent l'interrubain comme un luxe par rapport aux autres moyens de communication et aux autres dépenses de consommation. En conséquence, lorsqu'on leur demande d'estimer leur utilisation de ce service, comme dans une enquête, les répondants de ces groupes de revenu ont tendance à faire une sous-estimation parce qu'ils croient qu'ils devraient employer leur argent à autre chose. Pour ce qui est de la consommation réelle, cependant, les valeurs relatives peuvent changer puisque l'urgence d'une situation peut faire en sorte qu'un appel interrubain est la solution la moins coûteuse par rapport à d'autres moyens de communication. Ainsi, les

n'est pas statistiquement significatif ($p = .9905$) pour expliquer l'écart entre les valeurs de l'enquête préliminaire et celles de l'enquête-journal. Un test t portant sur un échantillon pour chacune des moyennes des quatre groupes a révélé que chaque moyenne était statistiquement différente de zéro au niveau de signification de 0.01. Par conséquent, les résultats supposent, en ce qui concerne chacun des quatre groupes d'utilisation, que les valeurs positives de la moyenne signifient que les répondants ont surestimé l'utilisation dans l'enquête préliminaire.

5.2 Liens des écarts entre les valeurs de l'enquête préliminaire et celles de l'enquête-journal avec certaines variables démographiques

Dans le tableau 1, nous avons observé un écart important entre les valeurs de l'enquête préliminaire et celles du carnet pour les mêmes répondants sur une période de douze mois. Il serait intéressant d'examiner ce qui explique le biais dû à l'interprétation dans l'enquête préliminaire. À cette fin, certains facteurs démographiques sont considérés. Nous avons déterminé plusieurs niveaux pour chaque facteur et nous avons fait une analyse de la variance à un critère de classification pour expliquer les écarts. Les tableaux 3 à 7 montrent les résultats de ces analyses.

5.3 Sexe

Le tableau 3 montre le lien de l'écart entre les valeurs de l'enquête préliminaire et celles du carnet avec le sexe du répondant. Les valeurs p de l'analyse de la variance à un critère de classification sont statistiquement significatives au niveau de 0.05 ou moins pour 9 des 12 mois et au niveau de 0.01 pour la moyenne des douze mois. Par conséquent, les résultats indiquent que les hommes comme les femmes surestiment leur utilisation réelle des services de l'interruption téléphonique et que les femmes la surestiment à un degré plus élevé que les hommes.

Tableau 3

Résultats de l'analyse de la variance à un critère de classification illustrant l'écart entre les valeurs de l'enquête préliminaire et celles du carnet en fonction du sexe du répondant

Mois	Écart selon le sexe		Analyse de la variance (Valeur p)
	Homme	Femme	
Février	0.412	1.135	0.006*
Mars	-0.015	0.818	0.005*
Avril	0.379	1.201	0.002*
Mai	0.310	1.304	0.008*
Juin	0.562	1.205	0.016**
Juillet	0.376	1.008	0.018**
Août	0.395	0.987	0.031**
Septembre	0.927	1.225	0.258
Octobre	0.605	1.149	0.042**
Novembre	0.593	1.003	0.129
Décembre	-0.112	0.464	0.041**
Janvier	0.164	0.675	0.075**
Moyenne ^a	0.380	0.990	0.010*
n	617	911	

^a Moyenne des douze mois.
* Significative à un niveau de 0.01.
** Significative à un niveau de 0.05.

Tableau 2

Résultats de l'analyse de la variance à un critère de classification illustrant l'écart entre les valeurs de l'enquête préliminaire et celles du carnet en fonction du degré d'utilisation de l'interurbain (moyenne des 12 mois)					
Degré d'utilisation					
Élevé	Moyen	Faible	Non-utilisation	Valeur p	
Ecart moyen (enquête préliminaire-carnet)					
0.762	0.799	0.795	0.580	0.9905	n
316	605	547	45		

du tableau 1 indiquent le contraire. Le nombre d'appels en décembre 1977 est beaucoup plus élevé qu'en décembre 1978. Par conséquent, nous sommes forcés de conclure qu'il y a effectivement un écart significatif dû à la méthode de collecte des données. Dans notre étude, les répondants ont surestimé le nombre de leurs appels lors de l'enquête préliminaire, si l'on compare avec le nombre d'appels qu'ils ont consigné dans leur carnet.

Avant d'analyser le lien entre l'écart entre les valeurs de l'enquête préliminaire et celles de l'enquête journal et plusieurs variables démographiques, il importe d'évaluer le rôle de l'utilisation réelle dans l'explication de cet écart. Notre raisonnement pour ce test est que si l'écart entre les valeurs estimées dans l'enquête préliminaire et les valeurs consignées dans le carnet est dû au niveau absolu d'utilisation, une analyse plus approfondie se révélerait douteuse puisque l'expérience tendrait à fausser notre variable dépendante (McKenzie 1983). Par ailleurs, si l'on ne peut attribuer aucune signification statistique aux écarts entre les valeurs obtenues par les deux méthodes et les degrés absolus d'utilisation, alors l'analyse en fonction des variables démographiques serait d'une plus grande validité.

Le tableau 2 montre les résultats de l'analyse du lien de l'écart entre les valeurs de l'enquête préliminaire [VALEUR ESTIMÉE 1] et celles du carnet [VALEUR OBSERVÉE 2] avec le degré absolu d'utilisation [VALEUR ESTIMÉE 2]. McKenzie a étudié les biais de réponse et les biais dus à l'enregistrement que comporte la collecte d'informations sur les appels téléphoniques par la méthode du carnet. Il a observé que le taux de réponse variait selon l'utilisation et que les taux d'enregistrement des appels téléphoniques avaient tendance à diminuer avec l'utilisation. Par conséquent, les données concernant les appels téléphoniques recueillies par cette méthode risquent de comporter plusieurs biais. Notre étude porte sur l'écart entre les valeurs recueillies lors de l'enquête préliminaire et celles consignées dans le carnet et l'utilisation réelle par les consommateurs; nous examinons particulièrement les divergences entre la consommation « estimée » et « réelle » et le niveau (degré) d'utilisation. Bien que les deux méthodes comportent des biais dus à l'enregistrement, l'écart entre les valeurs des deux méthodes n'est pas lié à l'utilisation.

En outre, on peut examiner la validité de la VALEUR ESTIMÉE 2 comme variable de classification en corrélant cette mesure avec la VALEUR OBSERVÉE 2 et la VALEUR ESTIMÉE 1. Ces dernières ont d'abord été divisées en quatre degrés d'utilisation: élevé, moyen, faible et non-utilisation, au moyen de différents seuils. On a ensuite croisé la VALEUR ESTIMÉE 1 et ces deux mesures qualitatives. Des liens statistiquement significatifs ont été observés dans tous les cas.

Notre variable dépendante est l'écart entre les valeurs de l'enquête préliminaire [VALEUR ESTIMÉE 1] et celles de l'enquête-journal [VALEUR OBSERVÉE 2] et notre variable indépendante est le degré d'utilisation divisé en quatre niveaux: élevé, moyen, faible et non-utilisation [VALEUR ESTIMÉE 2]. Les résultats d'une analyse de la variance à un critère de classification utilisant la méthode des moindres carrés indiquent que le degré d'utilisation

5. RÉSULTATS

5.1 Utilisation moyenne déclarée lors de l'enquête préliminaire et de l'enquête journal

Le tableau I montre le nombre moyen d'appels interurbains tiré du carnet des répondants pour chacun des douze mois et le nombre d'appels interurbains pour un mois typique selon l'enquête préliminaire [VALEUR ESTIMÉE 1]. Il est intéressant de noter que l'estimation faite lors de l'enquête préliminaire est beaucoup plus élevée que l'utilisation réelle déclarée [VALEUR OBSERVÉE 2].

Les moyennes des valeurs du carnet révèlent la présence d'un caractère saisonnier dans l'utilisation. C'est en décembre 1978 qu'a été enregistré le plus grand nombre d'appels, soit 4.123. Bien que l'enquête préliminaire demandait aux répondants de déclarer le nombre de leurs appels pour un mois moyen ou typique, il est fort probable que les répondants ont choisi de faire leur déclaration d'après leurs appels de décembre 1977 puisque l'enquête a été menée en janvier 1978. Un test *t* portant sur un échantillon indique que la moyenne des écarts entre la valeur obtenue à l'enquête préliminaire et la valeur consignée dans le carnet pour décembre (0.235) est sensiblement différente de zéro (valeur $p = 0.001$). De même, les résultats du test *t* pour les onze autres moyennes sont statistiquement significatifs. Ces résultats indiquent que les répondants ont en fait surestimé leur utilisation dans l'enquête préliminaire, si l'on considère les valeurs consignées dans le carnet.

Il est possible que les appels exceptionnellement nombreux en décembre 1977 influent sur l'estimation du nombre d'appels dans l'enquête préliminaire. Si c'est le cas, on pourrait dire que les résultats de notre étude comportent un biais dû à un facteur saisonnier. À cet égard, les auteurs ont examiné l'écart entre l'utilisation estimée dans l'enquête préliminaire et celle consignée dans le carnet de décembre 1978. Le fait de comparer les mêmes mois d'une année à l'autre pourrait aider à éliminer le facteur saisonnier. Comme l'indique le tableau I, cet écart est statistiquement significatif. Cependant, il peut être attribuable aux méthodes de collecte différentes et à un facteur de tendance étant donné que la comparaison porte sur deux années. Si nous supposons une tendance positive de l'utilisation de l'interurbain dans le temps, l'utilisation observée en décembre 1978 devrait être supérieure à celle de décembre 1977. Or, les données

Tableau I

Nombre absolu moyen d'appels interurbains
et estimations de l'enquête préliminaire

Mois	Nombre absolu moyen d'appels
Février	3.516
Mars	3.878
Avril	3.486
Mai	3.610
Juin	3.414
Juillet	3.604
Août	3.606
Septembre	3.250
Octobre	3.426
Novembre	3.518
Décembre	4.123
Janvier	3.891
Estimation de l'enquête préliminaire	4.358
<i>n</i>	= 1530

3. L'ETUDE

En 1978 et 1979, la société AT & T (American Telephone and Telegraph Company) a entrepris une importante collecte de données en vue de planifier et d'élaborer des stratégies commerciales pour son marché des communications téléphoniques interurbaines résidentielles. On a constitué un échantillon d'environ 4000 ménages, dont les caractéristiques pouvaient être généralisables à l'ensemble du pays, et on a demandé à ces personnes de faire partie d'un panel pendant douze mois. L'échantillon était démographiquement équilibré selon six variables: la densité de population, le revenu, l'état matrimonial, l'âge, le sexe et le lieu de résidence. Tous les membres du panel ont participé à une enquête préliminaire en répondant à un questionnaire envoyé par la poste, en janvier 1978. Une fois ce questionnaire rempli, chaque membre du groupe devait remplir un carnet hebdomadaire pendant les douze mois suivants. Au cours de l'étape préliminaire, on avait posé aux répondants la question suivante: « Dans un mois moyen ou typique, combien de fois communiquez-vous, pour des raisons autres que les affaires, avec des parents et des amis qui habitent à au moins 50 milles de chez vous? ». Nous appellerons ci-après cette mesure VALBUR ESTIMÉ 1. De plus, chaque membre du panel, lors de l'étape préliminaire, a répondu à la question suivante: « Quel est votre degré d'utilisation de l'inturbain: élevé, moyen, faible ou non-utilisation? ». Cette mesure est appelée ci-après VALBUR OBSERVÉ 1. De plus, l'ordre des catégories de réponses avait été choisi au hasard afin d'éviter tout risque de biais dû à l'ordre de classement. Après les douze mois, un échantillon de 2,350 répondants a été conservé. On s'est aperçu que l'érosion du panel pouvait être un problème éventuel dans cette étude étant donné que les taux d'érosion peuvent varier considérablement parmi des sous-groupes définis selon des critères démographiques. Afin de résoudre ce problème, un programme d'équilibrage de l'échantillon a été élaboré et utilisé pour permettre le choix au hasard, à partir du groupe des 2,350 répondants, d'un sous-échantillon de participants démographiquement équilibré. Après le contrôle et l'équilibrage de l'échantillon, 1,530 membres du panel qui avaient répondu au questionnaire préliminaire et rempli leur carnet pendant les douze mois ont été choisis pour cette étude.

4. ANALYSE DES DONNÉES

Dans une question importante du questionnaire préliminaire, on demandait aux répondants d'« estimer » leur utilisation pour un mois typique [VALBUR ESTIMÉ 1]. Afin d'obtenir une unité de mesure uniforme, nous avons regroupé les renseignements consignés dans le carnet hebdomadaire [VALBUR OBSERVÉ 1] en douze totaux mensuels pour chaque répondant. Ces valeurs seront appelées VALBURS OBSERVÉS 2. Les écarts entre l'estimation de l'utilisation déclarée dans le questionnaire de l'enquête préliminaire [VALBUR ESTIMÉ 1] et l'utilisation réelle déclarée dans le carnet [VALBUR OBSERVÉ 2] ont été calculés pour chaque répondant pour douze périodes d'un mois ainsi que pour la moyenne des douze mois. Nous avons fait, pour chaque mois et pour la moyenne des douze mois, une analyse de la variance à un critère de classification afin de déterminer s'il existait des écarts importants en fonction des niveaux de plusieurs variables démographiques: sexe, revenu, niveau d'instruction et âge. Nous avons ensuite fait un test de contraste a posteriori pour comparer toutes les paires possibles des moyennes de niveaux pour chaque variable démographique. Finalement, afin d'évaluer les effets des interactions parmi les quatre variables démographiques, nous avons fait une analyse de la variance à quatre critères de classification pour la moyenne des douze mois.

a conclu que les biais n'étaient pas un grave problème dans les enquêtes par panel. Sudman (1964a, b), lui, a constaté que l'abandon est plus fréquent chez les répondants de sexe masculin. En outre, il semble qu'il n'y ait aucun lien entre l'effort à consentir pour déclarer des renseignements et l'exactitude de ceux-ci ou le taux d'abandon chez les répondants d'un panel. Pour ce qui est de la perte de précision, McKenzie (1983) affirme que plus la période pendant laquelle on fait appel à un panel est longue, plus le groupe diminue, tandis que Sandage (1956) a observé que le recours répété à un panel donné n'introduisait pas de biais dans la précision des données déclarées.

Parfitt (1967) prétend que les maîtresses de maison qui répondent à des enquêtes ne se rappellent avec précision que les achats récents de produits fréquemment utilisés. Par conséquent, l'inscription dans un carnet des achats passés offre une mesure plus sûre et plus exacte que l'enquête-mémoire. Dans les enquêtes, on demande habituellement aux répondants de déclarer les achats qu'ils ont faits pendant une longue période ou de faire mentalement la moyenne de leurs dépenses pour une semaine ou un mois typique. En conséquence, Parfitt (1967) conclut qu'il est fort probable que les répondants exagèrent le montant et la fréquence de leurs achats et simplifient à l'extrême la complexité du processus de décision en matière de dépenses.

Comme nous l'avons déjà mentionné, notre recherche porte sur l'exactitude relative des données sur les dépenses de consommation déclarées lors d'une enquête-mémoire et dans une enquête-journal. Quelques articles seulement traitent de cette question de façon empirique. Wind et Lerner (1979) comparent la validité des deux méthodes pour les dépenses de consommation. Leurs données sont tirées d'un échantillon de 450 maîtresses de maison composant un panel de consommateurs pour une enquête-journal de la Market Research Corporation of America (MRCA). Après avoir répondu à un questionnaire envoyé par la poste, les maîtresses de maison devaient tenir un registre de leurs achats de diverses marques de margarine pendant six mois. Pour les deux méthodes de déclaration, on a observé une divergence entre les réponses individuelles et les réponses globales. Quand on considère les données d'ensemble, le questionnaire et le carnet permettent l'un et l'autre de déterminer l'importance des différentes marques sur le marché. Cependant, des divergences appréciables ont été observées au niveau des données individuelles. En effet, les données obtenues par le questionnaire étaient moins exactes que celles consignées dans le carnet. Les auteurs attribuent cette inexactitude à l'ignorance, aux déficiences de mémoire, à la mauvaise conception du questionnaire, aux erreurs de déclaration, à la falsification et à une erreur systématique de l'enquêteur.

Stanton et Tucci (1982), après les travaux de Wind et Lerner (1979), ont constitué un échantillon de 7,945 participants à la National Food Consumption Survey (1977-1978). Des interviews sur place ont été faites pour recueillir des données sur les dépenses alimentaires des participants au cours des vingt-quatre heures précédentes. On a ensuite demandé aux participants d'inscrire dans un carnet leurs dépenses d'aliments et de boissons pendant les deux jours qui ont suivi l'interview. Les résultats ont montré, au niveau des données globales, que les interviews sur place ont produit des renseignements aussi exacts et sûrs que ceux consignés dans le carnet. Vu la nature des données, les auteurs n'ont pu vérifier l'exactitude relative des deux méthodes au niveau des données individuelles.

Les divergences évidentes dans les résultats signalés par Wind et Lerner (1979) et Stanton et Tucci (1982) peuvent être attribuées aux différentes périodes de référence pour lesquelles les consommateurs devaient déclarer leurs dépenses. Dans l'étude de Wind et Lerner, on avait demandé aux répondants de déclarer la marque qu'ils achetaient le plus souvent. Des questions de ce genre exigent une meilleure mémoire puisque la période visée est assez longue. Par contre, dans l'étude de Stanton et Tucci, la période de référence est limitée aux vingt-quatre heures précédant l'interview. Parfitt (1967) affirme que les répondants parviennent mieux à déclarer des achats récents. La conclusion de Stanton et Tucci n'est donc pas vraiment surprenante et, de plus, ne contredit pas les résultats de l'analyse de Wind et Lerner puisque la mémoire des répondants avait été excellente tant pour répondre au questionnaire que pour remplir le carnet.

cas et d'études empiriques dans lesquels il traite des avantages et des inconvénients relatifs des deux outils d'inscription des dépenses, mais sans comparer leur exactitude relative. En règle générale, l'enquête-mémoire offre des avantages économiques tout en ayant certains des désavantages de la méthode du carnet. En raison de contraintes de temps et de ressources, la plupart des chercheurs utilisent l'enquête-mémoire malgré les multiples problèmes de mesure qu'elle comporte.

On considère généralement que la méthode du carnet est plus avantageuse que l'enquête-mémoire, surtout parce que les personnes qui tiennent un carnet ont la possibilité d'y inscrire un événement peu de temps après qu'il s'est produit. Pour cette raison, Sudman et Ferber (1971) ont presque mis en doute la méthode de l'enquête-mémoire pour la collecte de données sur les dépenses et ont suggéré l'utilisation exclusive du carnet. Mais la méthode du carnet n'est pas elle non plus exempte de défauts. Les auteurs ont examiné des données fournies par des ménages de la région de Chicago en 1972 et constaté qu'il y avait sous-déclaration du nombre d'achats par la méthode de l'enquête-mémoire. Ils ont également observé que les répondants avaient eu de la difficulté à diviser leurs achats en catégories précises avec cette méthode.

Plusieurs auteurs affirment que la méthode du carnet convient uniquement pour certaines catégories de dépenses (Ferber, 1968; Grooteart, 1986; Wind et Lerner, 1979; Stanton et Tucci, 1982). Selon Pearl (1968), le carnet personnel est préférable en raison de la précision des données consignées. La méthode convient particulièrement pour l'achat d'articles chers, alors qu'on semble avoir tendance à déclarer moins souvent les dépenses peu élevées. Grooteart (1986) est du même avis et suggère que tous les membres admissibles du ménage tiennent un carnet pour ne pas oublier de déclarer des dépenses. Dans des études distinctes sur la déclaration des dépenses de produits alimentaires, Wind et Lerner (1979) et Stanton et Tucci (1982) donnent des arguments tendant à démontrer la supériorité des enquêtes par panel.

La conception du carnet pose des problèmes de collecte (Kemsley 1961; Kemsley et Nicholson 1960; Lewis 1948; Sudman 1964a, b; Sudman et Ferber 1971; Walsh 1977). Kemsley (1961) et Kemsley et Nicholson (1960) ont examiné des carnets de dépenses de consommateurs pour une période de trois semaines en 1953. Ils ont décelé des écarts importants dans la déclaration des dépenses au cours de la période selon le genre de dépense et la saison. Lewis (1948) a étudié l'exactitude de la déclaration hebdomadaire des dépenses alimentaires et vestimentaires par rapport à la déclaration mensuelle. L'auteur a observé qu'il y avait une diminution de 16% dans la déclaration mensuelle des dépenses comparativement à la déclaration hebdomadaire. Sudman (1964a) et Sudman et Ferber (1974) ont étudié d'autres méthodes de collecte de données sur les dépenses des consommateurs. Ils ont examiné l'importance de la rétribution, de la formation des répondants et de la méthode de déclaration. Dans les études qu'ils ont faites, ils ont constaté que la rétribution favorisait la collaboration des répondants et augmentait l'exactitude des renseignements et que la formation directe aidait les répondants à fournir des renseignements plus précis. La fréquence des achats et la conception des formules de déclaration étaient également des facteurs importants dans l'exactitude des données déclarées.

D'autres études ont été faites, portant plus précisément sur la collaboration des unités de consommation (Kemsley et Nicholson 1960; Pearl 1968; Sudman et Ferber 1974). Kemsley et Nicholson (1960) affirment que l'importance de chaque achat influe considérablement sur le degré de collaboration des répondants. Pearl (1968) et Sudman et Ferber (1974) soutiennent que le montant et la durée de la rétribution suscitent la collaboration des répondants.

Un autre problème que pose la méthode du carnet, c'est l'érosion du panel par les abandons (Sandage 1956; Soder 1959; Sudman 1964a, b) et la perte de précision (McKenzie 1983; Sandage 1956; Soder 1959; Sudman 1964a, b). Sandage (1956) a cherché à déterminer si les consommateurs d'un panel introduisaient des biais à la longue. En examinant les résultats de trois enquêtes distinctes réalisées auprès de ménages agricoles de l'Indiana entre 1947 et 1954, l'auteur

Rôle des facteurs démographiques dans l'analyse de la précision de la déclaration des dépenses de consommation dans l'enquête-mémoire et dans l'enquête-journal

EDWARD R. BRUNING et MICHAEL Y. HU¹

RÉSUMÉ

Dans le présent article, les auteurs évaluent l'efficacité relative des méthodes de collecte de données que sont l'enquête-mémoire et l'enquête-journal dans le contexte du marché des communications téléphoniques interurbaines. Une analyse des réponses de 1,530 répondants montre que deux variables démographiques, le sexe et le revenu, expliquent la différence qui existe entre la déclaration des dépenses lors d'une enquête et l'inscription de ces renseignements dans un carnet.

MOTS CLÉS: Enquête; carnet; collecte de données.

1. INTRODUCTION

Une lecture attentive de la littérature spécialisée en marketing nous permet de constater notre manque de connaissances sur la précision relative des méthodes de l'enquête-mémoire et du journal pour recueillir des données sur les dépenses de consommation. Il est certain qu'une solution à ce problème serait utile aux chercheurs et aux personnes qui commandent une étude. Wind et Lerner (1979) soulignent l'importance d'évaluer adéquatement les deux méthodes et d'établir les caractéristiques des personnes qui déclarent fidèlement leurs achats par rapport aux personnes dont les dépenses déclarées diffèrent beaucoup des dépenses réelles. À cet égard, une analyse des différences porte surtout sur l'instrument de collecte des données puisque le choix de l'instrument pourrait influencer sur les décisions de gestion relatives au positionnement d'un produit, aux stratégies de segmentation du marché, aux médias publicitaires, à l'étude de textes publicitaires et aux tests de concept et de produit (Wind et Lerner 1979).

Le but de notre article est d'évaluer empiriquement les rapports entre plusieurs variables démographiques et les deux méthodes de déclaration des dépenses à partir d'un seul échantillon de répondants dans le marché américain des communications téléphoniques interurbaines. Nous apportons d'autres éléments au problème déjà soulevé par Wind et Lerner. D'abord, nous examinons l'état actuel de nos connaissances sur la nature des deux méthodes. Ensuite, nous décrivons les méthodes de recherche et présentons les résultats d'une enquête dans le marché des communications téléphoniques interurbaines. En conclusion, nous présentons certaines observations utiles aux fournisseurs et aux utilisateurs de données sur les dépenses de consommation.

2. REVUE DE LA DOCUMENTATION

Les deux principales méthodes d'inscription des dépenses de consommation des ménages sont l'enquête-mémoire, où l'on demande aux membres du ménage de se rappeler les dépenses faites au cours d'une période donnée, et l'enquête-journal, qui consiste à consigner dans un carnet quotidien ou hebdomadaire des dépenses précises. Neter (1970) donne des exemples de

¹ Edward R. Bruning et Michael Y. Hu, Graduate School of Management, Kent State University, Kent, Ohio, 44242, États-Unis.

DALENIUS, T. (1953). The multivariate sampling problem. *Skandinavisk Actuarietidskrift*, 36, 92-102.

DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almqvist and Wiksell.

FOLKS, J. L., et ANTLE, C. E. (1965). Optimum allocation of sampling units when there are R responses of interest. *Journal of the American Statistical Association*, 60, 225-233.

FORSYTH, G. E. (1968). On the asymptotic directions of the s -dimensional optimum gradient method. *Numerische Mathematik*, 11, 57-76.

HARTLEY, H. O. (1965). Multiple purpose optimum allocation in stratified sampling. *Proceedings of the Social Statistics Section, American Statistical Association*, 258-261.

HUDDLESTON, H. F., CLAYPOOL, P. L., et HOCKING, R. R. (1970). Optimum sample allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society A*, 139, 80-95.

KOKAN, A. R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society A*, 126, 557-565.

KOKAN, A. R., et KHAN, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society B*, 29, 115-125.

KUHN, H. W., et TUCKER, A. W. (1951). Nonlinear programming. *Proceedings 2nd Berkeley Symposium Mathematical Statistics and Probability*.

LUENBBERGER, D. G. (1984). *Linear and Nonlinear Programming*. Reading, Massachusetts: Addison-Wesley.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of representative sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, 558-625.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin and Company.

7. ANALYSE

Cet article nous a permis d'exposer de façon formelle la répartition optimale de l'échantillon dans les enquêtes à objectifs multiples avec des contraintes de variance linéaires et de définir des expressions pour les dérivées partielles de la fonction de coût par rapport aux contraintes de précision. Ces expressions, notamment, produisent des approximations utiles pour la planification d'enquêtes, permettant ainsi d'exécuter de nombreux travaux préliminaires sans les calculs rigoureux faits par ordinateur.

Les multiplicateurs de Lagrange normalisés α_j^* occupent une place importante dans cet article. Nous avons remarqué, en particulier, que lorsque la j -ième contrainte de variance n'est pas « agissante » dans la solution du problème de répartition, le j -ième multiplicateur de Lagrange $\alpha_j^* = 0$.

La méthode d'optimisation analysée dans cet article produit une solution continue qu'il faut arrondir d'une manière quelconque si l'on veut obtenir des tailles de strates en nombres entiers. De toute évidence, cette opération d'arrondissement créera un certain écart par rapport à l'optimalité. Toutefois, on s'entend normalement pour dire que la fonction objectif considérée ici est plutôt insensible à de faibles entorses à l'optimalité (voir Cochran 1977), de sorte que les solutions en nombres entiers sont probablement rentables. De fait, les erreurs d'arrondissement seraient vraisemblablement négligeables par rapport aux erreurs d'échantillonnage qui touchent les estimations de moyennes et de variances dont on dispose normalement pour élaborer un plan de sondage optimal.

Enfin, rappelons-nous que cet article ne tenait aucunement compte des facteurs de correction pour population finie. Nous pourrions facilement les intégrer au modèle de répartition en arrangeant les équations (1) et (3), mais cela rendrait l'équation (13) quelque peu imprécise. Néanmoins, il convient de se rappeler que même lorsque le facteur de correction pour population finie n'est pas négligeable pour quelques-unes des strates, l'effet global est, habituellement négligeable. Quoi qu'il en soit, il est toujours possible de calculer le terme perturbateur $\sum_{i=1}^I W_i^2 S_{ij}^2 / N_i$ pour une évaluation de la situation et de l'ajouter, si nécessaire, à v_k dans la formule (13) pour obtenir des résultats exacts.

REMERCIEMENTS

L'auteur a apprécié les commentaires des arbitres et les échanges fructueux qu'il a eus avec Rick Williams du Research Triangle Institute et Patrick McCarthy d'Applied Management Sciences; ces personnes ont contribué à améliorer sensiblement le contenu de cet article.

BIBLIOGRAPHIE

ARTHANARI, T.S., et DODGE, Y., (1981), *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

BETHEL, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Section, American Statistical Association*, 209-212.

CHATTERJEE, S. (1968). Multivariate stratified surveys. *Journal of the American Statistical Association*, 63, 530-534.

CHATTERJEE, S. (1972). A study of optimum allocation in multivariate stratified surveys. *Skandinavisisk Aktuarietidskrift*, 55, 73-80.

CHROMY, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Section, American Statistical Association*, 194-199.

Tableau 2.
Exemple de répartition de l'échantillon: solution optimale
pour un premier passage en machine.

Strate	$\sum \alpha_j^* a_{ji}$	x_i^*	Répartition optimale
Total:			241
1	33.6749	.0111	90
2	3.4495	.0347	29
3	2.9783	.0373	27
4	7.6294	.0233	43
5	4.9119	.0291	34
6	1.3554	.0553	18
Total: 241			
Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment chauffée au mazout
Multiplicateur de Lagrange (normalisé):	.3340	.0000	.0000
Niveau de précision atteint:	.0600	.0481	.0502
Prix fictifs à 10 %:	-32	-16	0

Tableau 3.

Exemple de répartition de l'échantillon: solution optimale
pour un échantillon dont la taille ne doit pas excéder 200.

Strate	$\sum \alpha_j^* a_{ji}$	x_i^*	Répartition optimale
Total:			201
1	33.6749	.0134	75
2	3.4495	.0418	24
3	2.9783	.0449	22
4	7.6294	.0281	36
5	4.9119	.0351	29
6	1.3554	.0666	15
Total: 201			
Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment chauffée au mazout
Multiplicateur de Lagrange (normalisé):	.3340	.0000	.0000
Niveau de précision atteint:	.0657	.0528	.0551
Prix fictifs à 10 %:	-27	-13	0

Exemple de répartition de l'échantillon : enquête sur les établissements d'enseignement.
Tableau 1.

Ecart type de strate					
Poids	Surface utile	Âge du bâtiment	Nombre d'employés	Proportion du bâtiment (%) chauffé au mazout	
Unités de précision normalisées					
Strate	1	22,319,11	43,71	25,72	48,15
	2	24,056,21	16,68	27,09	36,79
	3	54,201,75	24,70	17,11	48,04
	4	155,514,21	16,01	59,46	38,07
	5	125,239,21	14,74	51,27	48,80
	6	355,392,69	20,90	212,13	57,74
Moyenne:	V _k :	54,641,85	43,03	45,23	67,58
		.06	.06	.06	.06

Strate	Taille d'échantillon requise:			
	222	149	127	121
6	2.03	.01	1.06	.04
5	7.37	.01	.12	.05
4	11.36	.19	2.44	.45
3	3.83	1.28	.56	1.96
2	3.73	2.89	6.93	5.70
1	12.33	76.24	23.90	37.52

indique également les prix fictifs à 10 % : une hausse de 10 % de la première (ou de la seconde) contrainte entraînera une réduction de la taille de l'échantillon d'environ 32 (ou 16) unités. Comme les troisième et quatrième contraintes ne sont pas «agissantes» dans la solution, une modification des CV correspondantes n'aura aucun effet sur la répartition ou les coûts de l'échantillonnage.

Le tableau 3 donne une solution pour un second passage en machine moyennant que la taille de l'échantillon ne dépasse pas 200. Les solutions optimales sont donc multipliées par 241/200 (de sorte que chaque élément de la répartition optimale soit réduit par un facteur de 200/241) et les CV sont multipliées par $\sqrt{241/200}$. Les nouveaux prix fictifs à 10 % sont -27 et -13 pour les première et seconde contraintes respectivement, reflétant ainsi la diminution du coût global de l'enquête. Il convient de noter que les CV du tableau 3 sont environ 10 % plus élevées que celles du tableau 1, de sorte que la réduction de la taille de l'échantillon annoncée par les prix fictifs du tableau 2 (soit 48) se compare assez bien à la réduction réelle, qui est de 41 unités. (Les prix fictifs seront toujours quelque peu optimistes à cause de l'approximation linéaire).

considérablement selon la complexité du problème, le nombre de contraintes agissantes et, bien sûr, les caractéristiques du matériel. Les divers systèmes utilisés dans les circonstances (20 à 30 strates et 5 à 10 contraintes) ont été le Macintosh SE (temps d'exécution: de 30 secondes à 2 ou 3 minutes), le Leading Edge Model D (de 1 à 5 minutes), le Zilog System 8000 (de 5 à 60 secondes) et le Compaq précité (de 5 à 10 secondes). Néanmoins, les temps d'exécution sont généralement négligeables par rapport au temps nécessaire pour créer des fichiers et exécuter d'autres tâches préliminaires. L'algorithme SUMT, par exemple, peut prendre plusieurs heures pour trouver une valeur initiale acceptable. Un avantage notable de l'algorithme décrit ci-dessus en quatre étapes est qu'il ne nécessite pas de valeurs initiales externes. En outre, il est relativement facile à programmer, n'exigeant que 40 ou 50 lignes de programmation.

Chromy (1987) propose un algorithme encore plus simple. Nous pouvons adapter cet algorithme à notre notation et à notre approche générale de la façon suivante: Posons $\alpha_j^{(1)}$ $\equiv 1/J$, et, pour $n \geq 2$, Soit

$$\alpha_j^{(n)} = \alpha_j^{(n-1)} (a_j^j x^{(n-1)})^2 / \sum_j \alpha_j^{(n-1)} (a_j^j x^{(n-1)})^2 \quad 1 \leq j \leq J. \tag{19}$$

Comme l'algorithme décrit précédemment en quatre étapes, l'expression (19) ne nécessite pas de valeurs initiales externes; en outre, cet algorithme est encore plus facile à programmer et, selon plusieurs comparaisons, semble converger beaucoup plus rapidement. Malheureusement, il semblerait que cette convergence ne puisse se vérifier de façon formelle mais un nombre considérable d'applications (voir Chromy 1987, pour une analyse plus détaillée) permettent de croire que cet algorithme possède des propriétés de convergence intéressantes.

6. EXEMPLE

Les tableaux 1 à 3 présentent un exemple tiré d'une enquête sur les établissements commerciaux. (Seules les strates relatives aux établissements d'enseignement sont présentées ici.) Ces tableaux contiennent quatre des principales variables d'intérêt: surface utile, âge du bâtiment, nombre d'employés à plein temps et proportion (en pourcentage) dans laquelle le bâtiment est chauffé au mazout. Le tableau 1 renferme les données relatives à la variance de strate. Les unités de précision normalisées sont calculées en l'occurrence par la formule

$$a_{ij} = \frac{W^2 S^2_{ij}}{Y^2 v^2_j}$$

où $v_j = .06$ pour toutes les variables (de sorte que la demi-largeur d'un intervalle de confiance à 90 % équivalle à environ 10 % de la moyenne). Le tableau 1 donne également la taille d'échantillon requise pour une répartition de Neyman pour chacune des variables. On suppose que les coûts d'enquête sont les mêmes d'une strate à l'autre.

Le tableau 2 donne la solution pour un premier passage en machine; cette solution nécessite un échantillon de 241 unités. Dans le même tableau on trouve les coefficients de Lagrange normalisés et les niveaux de précision atteints; ces chiffres indiquent clairement que la surface utile et l'âge du bâtiment sont les variables prédominantes tandis que les autres variables ne sont pas «agissantes». Dans ce cas, on a utilisé la valeur initiale $\alpha^{(1)} = (1, 0, 0, 0)$; comme les troisième et quatrième contraintes étaient toujours satisfaites, il n'y a eu qu'une itération avec une recherche dichotomique en 9 étapes pour $t^{(1)}$. (Les estimations successives pour la valeur optimale ont été 1/2, 1/4, 3/8, 5/16, 11/32, 21/64, 43/128, 85/256 et 171/512.) Le tableau 2

$$= - \sum_{j=1}^I t \sum_{k=1}^I (\delta_{kj} - \alpha_j) \sqrt{c_k} \frac{2 \sqrt{c_I} \sum_{j=1}^I \alpha_j a_{Ij}}{\sqrt{c_I} \sum_{j=1}^I \alpha_j a_{Ij}} + O(t^2) \\ = ((t/2) \sqrt{g(x(\alpha))} (a_k^k x(\alpha) - 1) + O(t^2)).$$

Si nous acceptons que t tende vers zéro, alors il existe une valeur $t \in (0,1)$ pour laquelle

$$\sqrt{g(x(t\delta_k + (1-t)\alpha))} = h_{k\alpha}(t) > h_{k\alpha}(0) = \sqrt{g(x(\alpha))}$$

si et seulement si $a_k^k x(\alpha) > 1$. Nous pouvons donc conclure, d'après (15), que les contraintes sont satisfaites à la convergence; cette conclusion et l'expression (16) impliquent que $\lim_{n \rightarrow \infty} x(\alpha^{(n)}) = x^*$.

Dans l'exécution de l'algorithme, l'étape 2 nécessite la recherche d'une valeur $t^{(n)}$. Définissons $h_{k\alpha}(t)$ comme dans l'équation (17). Il est clair d'après l'analyse précédente que $a_k^k x(t\delta_k + (1-t)\alpha^{(n)}) = 1$ lorsque $h(t)$ (et par conséquent g) est à un maximum. De plus, comme $h_{k\alpha}(t)$ est strictement concave, $h_{k\alpha}(t)$ est non croissante en t et il y a donc un seul point où $h_{k\alpha}(t) = 0$. On peut donc exécuter une recherche dichotomique pour le point où est maximisée $h_{k\alpha}(t)$ en vérifiant simplement si $a_k^k x(t\delta_k + (1-t)\alpha^{(n)}) = 1$, ce qui est un moyen rapide d'obtenir une valeur approchée de $t^{(n)}$.

Nous avons vu plus haut que l'algorithme pose a_1 comme valeur initiale. Cela est tout à fait arbitraire puisque n'importe lequel des a_j , $1 \leq j \leq J$, conviendrait. En pratique, la contrainte pour laquelle la répartition optimale (c.-à-d. formule (5)) entraîne le coût le plus élevé est généralement un choix judicieux pour la valeur initiale.

Il convient de souligner que la seconde étape de l'algorithme nécessitera IJ calculs pour la formule (14) ainsi qu'une recherche dichotomique en $I0$ étapes par exemple, chacune d'elles comportant $3I + J + 1$ calculs pour la formule (15), tandis que la troisième étape nécessitera J calculs. Ainsi, chaque itération de l'algorithme est $O(IJ)$. D'après (18), à la n -ième itération $O(IJ)$.

$$h_{k\alpha}^k(0) \approx \frac{2}{1} h_{k\alpha}^k(0) (a_k^k x(\alpha^{(n)}) - 1)$$

de sorte que $a_k^k x(\alpha^{(n)})$ est approximativement proportionnelle à $h_{k\alpha}^k(0)$ (jusqu'à une constante additive) Au point de vue heuristique $h_{k\alpha}^k(0)$ est la «pente» de h dans la direction de a_k , ce qui donne à penser que l'algorithme est essentiellement une méthode du gradient (ou une méthode de la plus grande ascension en l'occurrence). Cela suppose, en retour, un taux de convergence linéaire (voir, par exemple, Forsyth 1968).

D'après notre expérience (voir Bethel 1985), l'algorithme converge rapidement pour la plupart des problèmes de complexité moyenne. Par exemple, des problèmes de répartition d'échantillon où il y a de 20 à 30 strates et de 5 à 10 contraintes ont été résolus en 3 à 5 secondes à l'aide de l'algorithme (sur un ordinateur Compaq 386/20 avec un co-processeur arithmétique 30387) alors qu'ils l'ont été en 6 à 8 secondes à l'aide d'une technique séquentielle de minimisation sans contrainte (séquential unconstrained minimization technique – SUMT), qui comprend un algorithme avec peine fondé sur la méthode du gradient. Le temps d'exécution varie

Pour un vecteur $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_J)$, définissons $\tilde{x}(\alpha)$ comme suit

$$\tilde{x}_i(\alpha) = \sqrt{c_i} / \left(\sqrt{\sum_{j=1}^J \alpha_j d_{ij}} \sum_{k=1}^K \sqrt{\sum_{j=1}^J c_k \alpha_j d_{kj}} \right) \text{ si } \sum_{j=1}^J \alpha_j d_{ij} > 0, 1 \leq i \leq I$$

$= \infty$ dans le cas contraire.

Notons que $\tilde{x}(\alpha^*) = x^*$. Or, l'algorithme itératif permettant de déterminer x^* est défini de la façon suivante:

1. Supposer $\alpha_j^{(1)} = \delta_{1j}$, $1 \leq j \leq J$.

2. À l'itération $n \geq 2$, trouver un indice k pour lequel

$$(a_k - a_j)' \tilde{x}(\alpha^{(n)}) \geq 0, 1 \leq j \leq J. \tag{14}$$

L'expression ci-dessus représente la contrainte que la solution optimale courante satisfait le plus difficilement. Si $a_k' \tilde{x}(\alpha^{(n)}) \leq 1$, on doit mettre fin à l'algorithme, sinon on doit trouver une valeur $t^{(n)} \in (0, 1)$ pour laquelle

$$g(\tilde{x}(t^{(n)} \delta_k + (1 - t^{(n)}) \alpha^{(n)}) \geq g(\tilde{x}(t \delta_k + (1 - t) \alpha^{(n)})) \text{ pour tous } t \in [0, 1]. \tag{15}$$

3. Supposer $\alpha_j^{(n+1)} = t^{(n)} \delta_{kj} + (1 - t^{(n)}) \alpha_j^{(n)}$.

4. Cesser l'itération lorsque $|\alpha_j^{(n+1)} - \alpha_j^{(n)}| < \epsilon$, $1 \leq j \leq J$, où ϵ est un critère de convergence préalable.

Afin de vérifier la convergence de l'algorithme, il convient tout d'abord de noter que $\tilde{x}(\alpha)$ minimise $g(x)$ à la condition que $\sum_{j=1}^J \alpha_j a_j' x \leq 1$. Par conséquent, comme $\sum_{j=1}^J \alpha_j a_j' x^* \leq \sum_{j=1}^J \alpha_j = 1$,

$$0 \leq g(\tilde{x}(\alpha^{(n)})) \leq g(x^*) \tag{16}$$

pour tous n . Par ailleurs, d'après (15), $g(\tilde{x}(\alpha^{(n)}))$ est non décroissante, ce $g(\tilde{x}(\alpha^{(n)}))$ qui implique qu'elle est convergente. Afin de montrer que $\tilde{x}(\alpha^{(n)}) \rightarrow x^*$, définissons tout d'abord

$$h^{k\alpha}(t) = \sqrt{\sum_{j=1}^I c_j \left(t \delta_{kj} + (1 - t) \alpha_j \right) d_{ij}} = \sqrt{g(\tilde{x}(t \delta_k + (1 - t) \alpha))}. \tag{17}$$

Puisque $h^{k\alpha}(t)$ est concave (c.-à-d., $-h^{k\alpha}(t)$ est convexe),

$$h^{k\alpha}(t) - h^{k\alpha}(0) = t h'(0) + O(t^2) \tag{18}$$

Par conséquent

$$\frac{\partial g(x^*)}{\partial v_k} = 2 \left(\sum_{i=1}^I \sqrt{c_i} \sum_{j=1}^J \alpha_j^* W_2^i S_{ij}^2 / v_j^2 \right) \frac{\sqrt{c_i} \sum_{j=1}^J \alpha_j^* W_2^i S_{ij}^2 / v_j^2}{-c_i \alpha_k^* W_2^i S_{ik}^2 / v_k^3}$$

(11)

$$= -2 \frac{\alpha_k^* \sqrt{g(x^*)}}{\sum_{i=1}^I \frac{a_{ik} \sqrt{c_i}}{\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}}}}$$

$$= -2 \frac{\alpha_k^* \sqrt{g(x^*)}}{\sum_{i=1}^I \frac{a_{ik} \sqrt{c_i}}{\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}}}} \left(\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}} \sum_{k=1}^K \sqrt{c_k} \sum_{j=1}^J \alpha_j^* a_{kj} \right)$$

$$= -2 \frac{\alpha_k^*}{v_k} g(x^*) a_k^* x^*.$$

L'équation (7) implique nécessairement que $\alpha_k^* = 0$ lorsque $a_k^* x^* < 1$, de sorte que

$$\frac{\partial g(x^*)}{\partial v_k} = -2 \frac{\alpha_k^*}{v_k} g(x^*).$$

(12)

Cette formule est un peu plus compliquée que l'expression utilisée habituellement pour les prix fictifs (voir, par exemple, Luenberger 1984) à cause de la relation complexe entre g et v_j .
Supposons maintenant que nous augmentons v_k de $(100\pi) \%$, $0 \leq \pi \leq 1$. Nous allons représenter par $x^* + \Delta x^*$ la perturbation résultante dans x^* . Selon (12),

$$g(x^* + \Delta x^*) - g(x^*) \approx \pi v_k \frac{\partial g(x^*)}{\partial v_k} = -2 \pi \alpha_k^* g(x^*).$$

(13)

Par conséquent, une augmentation de $(100\pi) \%$ de la k -ième contrainte de variance entraînera une réduction de $(100)(2\pi\alpha_k^*) \%$ du coût global de l'enquête.

5. CONSIDÉRATIONS RELATIVES À LA PROGRAMMATION

Dans cette section, nous examinons certains aspects techniques du calcul de x^* et présentons un algorithme simple pour déterminer x^* et les coefficients α_j^* en examinant les moyennes pondérées $\sum_{j=1}^J \alpha_j a_j$. Définissons δ_{ij}

$$\delta_{ij} = 1 \text{ si } i = j$$

$$= 0 \text{ si } i \neq j.$$

Par convexité, $g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*)$ (pour tous $x > 0$, $x^* > 0$). Alors, d'après

(8)

$$g(x) - g(x^*) \geq (x - x^*)' \nabla g(x^*) \geq 0.$$

Il s'ensuit que x^* est le minimum de $g(x)$ à la condition que

$$\sum_j \lambda_j a_j x \leq \sum_j \lambda_j \quad \text{pour tous } x > 0.$$

Puisque la minimisation de g n'est pas influencée par des constantes multiplicatives positives, x^* minimise aussi $g(x)$ pourvu que $\sum_{j=1}^J \alpha_j^* a_j x \leq 1$ et $x > 0$, où $\alpha_j^* = \lambda_j / \sum_{j=1}^J \lambda_j$. Si nous voulons maintenant utiliser la formule (5) pour la répartition optimale dans les enquêtes à plusieurs variables, il suffit de l'appliquer à la somme pondérée $\sum_{j=1}^J \alpha_j^* a_j$:

$$x_i^* = \sqrt{c_i} / \left(\sqrt{\sum_{j=1}^J \alpha_j^* a_{ij}} \sum_{k=1}^K \sqrt{c_k} \sum_{j=1}^J \alpha_j^* a_{kj} \right) \quad \text{si } \sum_{j=1}^J \alpha_j^* a_{ij} > 0, \quad 1 \leq i \leq I \quad (9)$$

dans le cas contraire.

Donc, comme x^* minimise $g(x)$ à la condition que $a_j^* x \leq 1$, $x > 0$ pour $1 \leq j \leq J$, mx^* minimisera $g(mx)$ pourvu que $a_j^*(mx) \leq m$, $x > 0$ pour $1 \leq j \leq J$. Ainsi, comme nous l'avons indiqué plus haut, les contraintes de variance (ou CV) peuvent être rajustées par un facteur m (ou \sqrt{m}) si les coûts de l'enquête sont trop élevés. Evidemment, la formule (9) n'est mathématiquement utile que si les α_j^* sont connus. Néanmoins, elle est utile pour calculer les prix fictifs et mettre au point un algorithme permettant de déterminer x^* et les α_j^* .

4. SENSIBILITÉ DU COÛT DE L'ENQUÊTE AUX CONTRAINTES DE VARIANCE

Dans beaucoup de problèmes d'optimisation, il est utile de savoir ce que devient la solution optimale lorsque les contraintes sont modifiées légèrement. Cela peut être particulièrement le cas dans la conception d'enquêtes, où il faut souvent trouver le juste équilibre entre les coûts, les opérations d'enquête et le niveau de précision recherché. Quoi qu'il en soit, les «prix fictifs», définis par $\partial g(x^*) / \partial v_k$, sont utiles pour discerner les faibles variations des contraintes de variance, qui pourraient réduire sensiblement le coût global de l'enquête. En combinant les équations (2), (3), et (9), nous constatons facilement que le coût de la répartition optimale est

$$g(x^*) = \left(\sum_{i=1}^I \sqrt{c_i} \sum_{j=1}^J \alpha_j^* a_{ij} \right)^2 = \left(\sum_{i=1}^I \sqrt{c_i} \sum_{j=1}^J \alpha_j^* W_j^i S_{ij}^2 / v_j^2} \right)^2. \quad (10)$$

que nous appellerons «unités de précision normalisées». Notons que $a_{ij} \geq 0$. En nous servant de cette notation, nous pouvons exprimer le problème de la répartition optimale de la façon suivante:

Minimiser $g(x)$

à la condition que $a_j x \leq 1, \quad j = 1, 2, \dots, J$

(4) $x > 0$

où a_j est le j -ième vecteur colonne de la matrice $A = \{a_{ij}\}$.

Kokan (1963) analyse en profondeur ce modèle de répartition et montre comment on peut l'adapter à de nombreux problèmes courants de répartition de l'échantillon, notamment l'échantillonnage en grappes et l'échantillonnage double. Kokan et Khan (1967) poussent plus loin l'analyse dans ce contexte; Arthanari et Dodge (1981) exposent à nouveau les résultats de Kokan et Khan. Dans la même ligne de pensée, Kish (1976) décrit une catégorie de «formes linéaires» qui reviennent souvent dans la conception d'enquêtes et auxquelles s'appliqueront une bonne partie des résultats obtenus ici.

3. RÉPARTITION OPTIMALE

Le modèle de répartition optimale pour une variable unique est bien connu. Dans ce cas $J = 1$, et le minimum de $g(x)$ pourvu que $a_1 x \leq 1$ où $x > 0$, désigné par x^* , est défini

$$x_1^* = \sqrt{c_1 / \left(\sum_{k=1}^K \sqrt{c_k a_k} \right)} \quad \text{si } a_1 > 0, \quad 1 \leq i \leq I$$

(5) $= \infty$ dans le cas contraire.

Dans cette section, nous allons généraliser la formule (5) pour les cas où $J > 1$. La fonction g définie en (2) est strictement convexe pour $x > 0$, et les contraintes indiquées en (4) sont linéaires, de sorte que les résultats de base de la programmation convexe s'appliquent ici sans problème. Kokan et Khan (1967) ont démontré qu'il existe toujours une solution optimale. Comme ci-dessus, désignons la solution optimale par x^* . Alors, d'après le théorème de Kuhn-Tucker (1951), il existe des $\lambda_j \geq 0$ de telle sorte que

$$\Delta g(x^*) + \sum_j \lambda_j a_j = 0$$

(6)

(Δ désigne le gradient) et

$$\lambda_j \left(a_j x^* - 1 \right) = 0$$

(7)

pour $j = 1, 2, \dots, J$. Si $x > 0$ satisfait $\sum_{j=1}^J \lambda_j a_j x \leq \sum_{j=1}^J \lambda_j$, alors en combinant (6) et (7), nous avons

$$-x' \Delta g(x^*) = \sum_{j=1}^J \lambda_j a_j' x \leq \sum_{j=1}^J \lambda_j = \sum_{j=1}^J \lambda_j a_j' x^* = -x^{*'} \Delta g(x^*).$$

(8)

qui compensent largement les inconvénients de la méthode de la «programmation convexe». La première constatation est qu'une réduction proportionnelle de la répartition optimale (dans les enquêtes à plusieurs variables) donne une répartition qui est optimale suivant des contraintes qui sont proportionnelles aux contraintes initiales. En conséquence, si la solution optimale est trop coûteuse, elle peut être ramenée directement à un niveau plus conforme au montant budgété et on peut évaluer directement les effets de cette opération sur la précision des estimations échantillonnelles. La seconde constatation est que l'on obtient une expression simple pour les dérivées partielles du coût de la répartition de l'échantillon par rapport aux contraintes de variance. Ces quantités, que l'on appelle prix fictifs, illustrent la sensibilité du coût aux contraintes de variance et sont utiles pour l'évaluation du rapport coût/efficacité du plan de sondage. Toutefois, l'exécution même de l'optimisation convexe demeure un problème. Beaucoup d'articles ont été écrits sur des méthodes permettant de résoudre les problèmes de programmation de ce genre et il existe de nombreux logiciels conçus à cette fin. Néanmoins, nous allons faire ici des considérations spéciales à ce propos et nous allons exposer une méthode de résolution simple. Cet algorithme, qui est essentiellement une méthode du gradient, est convergent, facile à programmer, et facile à utiliser puisqu'il ne requiert pas de valeur initiale. Un exemple nous permettra d'exposer cet algorithme et les autres méthodes mentionnées ci-dessus.

2. LE MODÈLE DE RÉPARTITION

Considérons le cas d'un échantillonnage aléatoire stratifié avec I strates et J variables. Supposons que la j -ième variable doit satisfaire l'équation

$$(1) \qquad \qquad \qquad \text{Var}(\bar{y}_j) \approx \sum_I^I W_i^2 S_{ij}^2/n_i \leq v_j^2,$$

où S_{ij}^2 , n_i , et W_i^2 désignent respectivement la variance de la j -ième variable de réponse, la répartition de la taille de l'échantillon et la proportion de la population qui se trouve dans la i -ième strate, et où v_j est une constante positive arbitraire. Nous supposons dans cet article que les facteurs de correction pour population finie sont négligeables. En pratique, les effets de cette hypothèse, dont nous discuterons plus longuement dans la section 7, devraient être limités.

Soit

$$x_i = 1/n_i \text{ si } n_i \geq 1$$

$$= \infty \qquad \text{dans le cas contraire}$$

et supposons la fonction de coût

$$(2) \qquad \qquad \qquad g(x) = \sum_I^{I=1} c_i/x_i, \quad c_i > 0, \quad i = 1, 2, \dots, I.$$

Nous pourrions inclure dans cette expression un terme constant pour les coûts fixes, mais cela ne changerait rien au processus de minimisation; par conséquent, nous omettrons ce terme afin de simplifier la notation. Définissons les constantes

$$(3) \qquad \qquad \qquad a_{ij} = w_i^2 S_{ij}^2/v_j^2$$

Répartition de l'échantillon dans les enquêtes à plusieurs variables

JAMES BETHEL¹

RÉSUMÉ

Dans les enquêtes à objectifs multiples, on réalise souvent la répartition optimale de l'échantillon en définissant des contraintes linéaires pour la variance puis en utilisant la programmation convexe pour minimiser le coût de l'enquête. En nous servant du théorème de Kuhn-Tucker, nous définissons dans cet article une formule de répartition optimale en fonction des multiplicateurs de Lagrange. Cette formule sert ensuite à déterminer la dérivée partielle de la fonction de coût par rapport à la k -ième contrainte de variance; cette dérivée partielle est $-2\alpha_k^* g(x^*)/v_k$, où $g(x^*)$ est le coût de la répartition optimale et α_k^* et v_k sont respectivement, le k -ième multiplicateur de Lagrange normalisé et la borne supérieure pour la précision de la k -ième variable. Nous illustrons l'application de ces résultats à un plan de sondage à l'aide des données d'une enquête sur les établissements commerciaux.

MOTS CLÉS : Répartition de l'échantillon avec objectifs multiples; programmation non linéaire; échantillonnage stratifié.

1. INTRODUCTION

La répartition optimale de l'échantillon dans les enquêtes à objectifs multiples est un sujet qui a été traité pour la première fois par Neyman (1934), lorsque celui-ci élaborait sa théorie sur la répartition optimale dans les enquêtes à une variable. Depuis lors, de nombreux statisticiens se sont penchés sur la question et ont proposé plusieurs méthodes qui, pour la plupart, peuvent être classées en deux grandes catégories. La première catégorie consiste à calculer la moyenne pondérée des variances de strates et à trouver la répartition optimale pour la «variance moyenne» ainsi calculée. Dalenius (1953), Yates (1960), Folks et Antle (1965), Hartley (1965) et Kish (1976) analysent des méthodes de ce genre. La seconde catégorie de méthodes consiste à imposer une contrainte à chaque variance (sous forme d'une inéquation) puis à utiliser la programmation convexe pour déterminer le mode de répartition optimale qui satisfait toutes les contraintes. Dalenius (1957), Yates (1960), Kokan (1963), Hartley (1965), Kokan et Khan (1967), Chatterjee (1969, 1972), Huddleston, Claypool et Hocking (1970), Bethel (1985) et Chromy (1987) étudient tous l'un après l'autre l'utilisation de la programmation convexe par rapport à la répartition optimale dans les enquêtes à plusieurs variables. Chaque catégorie de méthodes a ses avantages et ses inconvénients. La méthode de la «moyenne pondérée» est simple au point de vue du calcul, intéressante sur le plan intuitif et peut être appliquée suivant une hypothèse de coût fixe, mais les poids sont choisis de façon arbitraire et les propriétés d'optimalité ne sont pas clairement définies. La méthode de la «programmation convexe» apporte une solution optimale au problème posé, mais à un coût qui peut sembler exorbitant, de sorte qu'il sera nécessaire de chercher une solution optimale qui tienne compte des considérations budgétaires.

Dans cet article, nous allons définir une expression en forme analytique pour la répartition optimale en fonction des multiplicateurs de Lagrange; cette expression est soumise à des contraintes linéaires (sous forme d'inéquations). Cela va nous permettre de faire deux constatations

¹ James Bethel, Westat, 1650 Research Boulevard, Rockville, MD, 20850 USA.

- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Sér. B*, 12, 241-255.
- PEARLMAN, J.G. (1980). An algorithm for the exact likelihood of a high-order autoregressive-moving average process. *Biometrika*, 67, 232-233.
- RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F., et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistics Review*, 45, 13-28.
- SINGH, D. (1968). Estimates in successive sampling using multi-stage design. *Journal of the American Statistical Association*, 63, 99-112.
- SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. Dans *Survey Sampling and Measurement*, (éd. N.K. Namboodini), New York: Academic Press, 201-216.
- STATISTIQUE CANADA (1985). Canadian Travel Survey: Estimation and variance estimation procedures. Rapport technique, Statistique Canada.
- TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.
- TIKKIWAL, B.D. (1979). Successive sampling — a review. *Bulletin of the International Statistics Institute*, 48, 367-384.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- YOUNG, P. (1984). *Recursive Estimation and Time Series Analysis: An Introduction*. New York: Springer-Verlag.

La simulation que nous avons faite donne à penser que même pour de petites séries de données, l'approximation asymptotique de la variance des estimations lissées est tout à fait acceptable. Cependant, lorsqu'il s'agit d'applications plus classiques de l'analyse chronologique, l'approximation asymptotique de l'erreur d'échantillonnage des valeurs estimées des paramètres peut laisser à désirer.

REMERCIEMENTS

Les auteurs tiennent à remercier un rédacteur associé de la revue et les arbitres pour leurs précieux commentaires sur des versions antérieures de cet article. Ils tiennent à exprimer plus particulièrement leur reconnaissance à l'arbitre qui, par ses commentaires profonds et éclairants, a contribué grandement à l'amélioration de cet article. Ils remercient également Pierre Hubert, chef de la Section des voyages, du tourisme et des loisirs, Division de l'éducation, de la culture et du tourisme, pour avoir mis à leur disposition les données de l'enquête sur les voyages des Canadiens. Des éléments de cet article sont tirés du mémoire de maîtrise présenté par le second auteur à l'Université de Guelph.

BIBLIOGRAPHIE

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in Time. Dans *Handbook of Statistics*, vol. 6, (éds, P.R. Krishnaiah et C.R. Rao), Amsterdam: Elsevier Science, 187-211.

BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Sér. B, 35, 61-68.

ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*. 26, 664-685.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 247-257.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33, 387-397.

HARVEY, A.C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3, 245-275.

HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.

JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, Sér. B, 42, 221-226.

JONES, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389-395.

KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82, 35-45.

MEINHOLD, R.J., et SINGPURWALLA, N.D. (1983). Understanding the Kalman Filter. *The American Statistician*, 37, 123-127.

MIYAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. Thèse de doctorat Iowa State University, Ames, Iowa.

série de données originales. Nous avons produit dans ces conditions cent séries de données pour chaque modèle. Les tableaux 1 et 2 donnent le biais (en pourcentage) des valeurs lissées ainsi que la précision (en pourcentage) relative à la différence entre les valeurs lissées et les valeurs réelles fondées sur ces simulations.

Afin de déterminer si 100 simulations étaient suffisantes pour estimer la précision, nous avons calculé une estimation du coefficient de variation de l'estimateur de la précision. Les simulations nous ont permis d'établir une estimation non biaisée de la variance de l'estimateur de l'erreur quadratique moyenne. Nous avons ensuite estimé la variance de l'estimateur de la précision par la méthode de linéarisation de Taylor. Les valeurs estimées des coefficients de variation allaient de 6 à 11 % pour les voyages faits sur le territoire de la Saskatchewan et de 5 à 9 % pour les voyages faits au Manitoba. Ainsi, ces estimations de la précision nous donnent une idée assez juste des conséquences de l'omission de l'erreur d'échantillonnage des paramètres autorégressifs.

Selon les tableaux 1 et 2, les biais obtenus par simulation sont tous faibles; de fait pour les deux séries de 22 observations chacune, on n'a relevé que quatre biais significatifs à un seuil de 5 % en utilisant un test *t*-ordinaire.

Nous remarquons aussi que les estimations de la précision obtenues par simulation (en pourcentage) sont en règle générale plus élevées que les valeurs figurant dans la colonne « C.V. des estimations lissées ». Cela est normal puisque les simulations tiennent compte des erreurs d'échantillonnage qui découlent de l'estimation de α et de θ^2 . Néanmoins, les C.V. des estimations lissées donnent une approximation acceptable des valeurs simulées, de sorte que nous pouvons dire que le fait de ne pas tenir compte de l'erreur d'échantillonnage de α et de θ^2 ne modifie pas réellement les coefficients de variation.

Les tableaux 3 et 4 contiennent des résultats de la simulation pour les paramètres estimés. En ce qui a trait aux coefficients de régression, un seul des biais simulés est significatif à un seuil de 5 %. Les erreurs types sont toutes conformes aux résultats de la simulation.

En revanche, les choses ne sont pas aussi simples en ce qui concerne les valeurs estimées de α et de σ^2 . Les biais des valeurs estimées de α sont fortement significatifs et comme nous pouvons le voir dans les tableaux 3 et 4, le biais d'une des valeurs estimées de σ^2 est aussi fortement significatif. Les erreurs types simulées ne sont pas très proches de l'approximation asymptotique de l'erreur type obtenue en inversant la matrice d'information de Fisher. Il semble que l'échantillon de notre exemple ne soit pas assez grand pour permettre une approximation asymptotique très précise. C'est un problème que l'on retrouve souvent lorsqu'on analyse de courtes séries chronologiques.

6. CONCLUSION

Lorsque la variance de l'erreur de sondage est faible par rapport à la variance de l'erreur de modèle, il y a peu de différence entre l'estimation lissée et l'estimation linéaire non biaisée à variance minimum et l'erreur type correspondante ne subit pas de réelle diminution même lorsque le modèle hypothétique est exact. Cependant, en ce qui a trait à l'estimation pour petits domaines, où l'erreur d'échantillonnage est élevée, l'erreur type de l'estimation lissée peut être beaucoup moindre que celle de l'estimation d'enquête. Par exemple, le lissage des estimations a donné des résultats plus notables dans le cas des voyages faits au Manitoba puisqu'il y avait une variance de l'erreur d'échantillonnage plus élevée dans ce cas-là.

Une des conséquences de la modélisation pour les enquêtes répétées est qu'une définition erronée du modèle peut se traduire par un estimateur à EQM minimum fortement biaisé. Il est donc essentiel de choisir un modèle qui soit conforme aux données et qui reflète une connaissance approfondie du phénomène en question. Comme notre exemple a trait à une petite série de données, un grand nombre de modèles statistiques seraient conformes aux données.

Tableau 4

1 Valeur estimées des paramètres pour les voyages-personnes de la Saskatchewan au Manitoba

Paramètre	Compte non tenu de l'erreur d'échantillonnage		Compte tenu de l'erreur d'échantillonnage	
	Valeur estimée	Valeur	Erreur type	Précision
			Biais	simulée
			simulé	
				Valeur t
				du biais

RÉGRESSION					
Ordonnée à l'origine (γ_0)	51.2	50.5	1.9	2.0	0.4
Termes linéaire (γ_1)	-0.17	-0.13	0.18	0.17	-0.04
1er trimestre (γ_2)	-20.1	-17.2	3.4	3.5	-0.6
2e trimestre (γ_3)	-5.9	-6.1	3.6	3.7	-0.1
3e trimestre (γ_4)	30.7	30.8	3.7	3.7	0.0
ARMA					
Autorégressif (α)	0.14	-0.75	0.66	0.71	0.49
Variance de modèle (σ^2)	100.0	5.7	18.7	9.5	-0.3
					7.90
					-0.29

1 Les valeurs simulées et les valeurs t sont fondées sur un échantillon de taille n = 100.

Les estimations lissées et les coefficients de variation correspondants figurent dans les tableaux 1 et 2. On calcule ces coefficients de variation en tenant compte de l'erreur d'échantillonnage des coefficients de régression, $\gamma_0, \dots, \gamma_4$. On peut procéder ainsi puisque, étant donné α et σ^2 , les estimations lissées sont une fonction linéaire des estimations originales, de sorte que les variances peuvent être calculées à l'aide de cette fonction linéaire et de la variance des résidus de régression du modèle hypothétique. Or, on n'a pas tenu compte jusqu'ici des erreurs d'échantillonnage pour les valeurs estimées de α et σ^2 . Nous voyons ci-dessous quels effets cela peut avoir.

Les estimations lissées pour le nombre de voyages-personnes en Saskatchewan sont généralement proches des estimations d'enquête, sauf peut-être en ce qui concerne l'été de 1980 et l'hiver de 1986. Les estimations lissées pour le nombre de voyages-personnes au Manitoba sont également proches des estimations d'enquête, sauf peut-être en ce qui concerne l'automne de 1980. Les trois exceptions peuvent être des valeurs aberrantes ou peuvent être attribuables à un événement spécial qui a contribué à accroître le tourisme durant ces trois trimestres. En règle générale, il y a deux façons d'inclure ces phénomènes dans le modèle: (i) soit accroître la variance de modèle dans le modèle d'espace d'états pour ces périodes ou prévoir des variables auxiliaires appropriées pour les événements spéciaux; (ii) soit accroître la variance d'échantillonnage pour les valeurs aberrantes. Seule une connaissance plus approfondie des circonstances nous permettrait de déterminer si ce sont là les ajustements qui conviennent. L'analyse que nous faisons ici peut nous permettre de faire ressortir les cas exceptionnels possibles.

Comme notre analyse n'a pas tenu compte jusqu'ici de l'effet de l'erreur d'échantillonnage liée à l'estimation de α et de σ^2 , nous avons effectué une simulation visant à évaluer cet effet. Jones (1979), Hamilton (1986) et Tam (1987) estiment qu'il ne faut pas négliger cette erreur d'échantillonnage, surtout lorsque la série chronologique comporte peu d'observations. Pour ce qui a trait à la simulation, nous avons produit des séries de données aléatoires suivant le modèle posé en (5.1) et (5.2). Nous nous sommes servi des estimations les plus vraisemblables comme valeurs des paramètres et du même schéma de non-réponse que pour la

Tableau 2
 Voyageurs-personnes d'une nuit ou plus sur le territoire du Manitoba -
 Résidents de la Saskatchewan¹

C.V. des C.V. des
 Nbre de Estimation brute Estimation lissée Estimation
 groupes de (milliers) (milliers) (milliers)
 Année Trimestre renouvellement
 Précision simulée (%)
 Biases simulées (%)

1979	Hiver	1	27	34	28.6	13.4	14.1	0.5
	Printemps	1	33	48	26.7	11.0	10.2	0.9
	Été	3	78	80	11.4	6.6	7.1	1.3
	Automne	3	55	48	12.9	10.1	10.8	0.6
1980	Hiver	1	24	30	29.7	13.6	14.5	0.5
	Printemps	3	63	50	12.3	9.5	9.4	0.7
	Été	1	86	80	19.0	6.6	6.3	0.8
	Automne	1	75	46	19.9	11.0	12.2	0.5
1981	Hiver	3	42	34	14.2	11.3	13.2	1.0
	Été	3	79	82	11.3	5.9	5.7	0.1
1982	Hiver	1	33	34	26.5	12.5	13.2	-2.8
	Printemps	1	46	44	23.7	10.7	10.0	1.6
	Été	3	78	82	11.4	5.7	5.4	0.1
	Automne	1	30	42	27.6	10.9	11.4	0.3
1984	Hiver	1	36	34	25.7	13.8	16.8	-1.3
	Printemps	1	48	43	23.4	11.4	11.5	0.1
	Été	3	82	82	11.1	6.1	7.3	-0.2
	Automne	1	30	40	27.7	11.5	11.4	0.6
1986	Hiver	1	33	33	26.7	16.3	19.9	-0.8
	Printemps	3	38	41	14.6	10.9	11.7	-0.1
	Été	3	90	81	10.8	7.1	8.8	-0.3
	Automne	3	42	40	14.1	11.2	10.5	1.7

¹ Il n'y a pas eu d'enquête sur les voyages des Canadiens au printemps et à l'automne de 1981 ni en 1983 et 1985. Les valeurs simulées figurant dans les deux dernières colonnes sont fondées sur un échantillon de taille = 100.

Tableau 3

Valeurs estimées des paramètres pour les voyageurs-personnes à l'intérieur de la Saskatchewan¹

Paramètre	Compte non tenu de l'erreur d'échantillonnage		Compte tenu de l'erreur d'échantillonnage		
	Valeur estimée	Valeur estimée	Erreur type	Précision simulée	Biases simulées
Valeur t	du biais	du biais	du biais	du biais	du biais

RÉGRESSION

Ordonnée à l'origine (γ_0)	831.4	815.0	15.6	14.4	1.8
1 ^{er} trimestre (γ_1)	-0.84	-0.86	1.52	1.51	-0.10
2 ^e trimestre (γ_2)	-209.6	-203.8	21.8	24.6	-3.5
3 ^e trimestre (γ_3)	-4.0	7.1	22.9	23.8	0.4
4 ^e trimestre (γ_4)	340.1	316.0	21.2	23.4	-0.4
Autocorrélatif (α)	0.14	0.47	0.66	0.68	-0.39
Variance de modèle (σ^2)	7930.5	879.3	1205.6	770.0	-488.2
					-6.77
					-8.16

¹ Les valeurs simulées et les valeurs t sont fondées sur un échantillon de taille n = 100.

l'EVC soit choisi de manière qu'il n'y ait pas de chevauchement des panels d'une fois à l'autre, l'hypothèse de l'indépendance n'est approximativement juste que lorsqu'il y a une faible corrélation des erreurs d'échantillonnage d'un trimestre à l'autre à l'intérieur de la même UPE. Cette hypothèse n'a pas été vérifiée. Les coefficients de variation (en pourcentage) ont été calculés au moyen de la fonction suivante:

$$CV = \alpha y_{-b} / \sqrt{\text{nombre de groupes de renouvellement}},$$

où y est l'estimation d'enquête en milliers. C'est la fonction que nous SCR recommandons d'utiliser pour les données de l'EVC concernant les résidents de la Saskatchewan; voir Statistique Canada (1985). Dans cette étude, nous nous servons d'un modèle de régression linéaire appliqué à des données de 1979 pour estimer la valeur des paramètres α et β les valeurs estimées sont respectivement 91.7528 et 0.353253. Pour les besoins de notre exemple, ces coefficients de variation ont été arrondis au dixième pourcent près.

Nous avons supposé le modèle:

$$(5.1) \quad y_t = \theta_t + e_t,$$

où les e_t 's sont des erreurs de sondage indépendantes, $e_t \sim N(0, s_t^2)$ et

$$(5.2) \quad \theta_t = \gamma_0 + \gamma_1 t + \gamma_2 Q_{1t} + \gamma_3 Q_{2t} + \gamma_4 Q_{3t} + e_t,$$

où $\{e_t\}$ suit un processus ARMA(1,0) avec des paramètres (α, σ^2) . Les termes de régression de l'équation (5.2) sont, dans l'ordre, l'ordonnée à l'origine, un terme indiquant le numéro du trimestre, où t varie linéairement en fonction du temps, son intervalle de variation étant [-15.5, 15.5], et enfin des termes saisonniers pour les trois premiers trimestres de chaque année, où

$$Q_{it} = 1 \text{ si la } i\text{-ième observation est dans le } i\text{-ième trimestre;}$$

$$= -1 \text{ si la } i\text{-ième observation est dans le quatrième trimestre;}$$

$$= 0 \text{ dans les autres cas;}$$

$$\text{pour } i = 1, 2, 3.$$

On pourrait trouver de meilleurs modèles pour ces données mais compte tenu de la faible dimension de la série, les tests d'hypothèses pour d'autres modèles ne seraient pas très puissants.

Pour calculer l'estimation la plus vraisemblable des paramètres inconnus de ce modèle, il faut tenir compte des hypothèses qui ont été posées à propos des erreurs de sondage. La plupart des utilisateurs de données officielles ne se préoccupent pas de l'erreur de sondage et supposent implicitement que les données d'entrée ne sont pas entachées d'erreur. Cela n'a pas de conséquence grave lorsque la variance de l'erreur de sondage est faible par rapport à la variance de l'erreur de modèle.

Les tableaux 1 et 2 donnent les estimations d'enquête et les coefficients de variation correspondants tandis que les tableaux 3 et 4 contiennent les résultats de l'estimation par la méthode du maximum de vraisemblance. Dans ce dernier cas, deux estimations sont indiquées pour chaque modèle: une pour le cas où on tient compte de l'erreur d'échantillonnage et une pour le cas où on n'en tient pas compte. On suppose que le modèle (5.2) s'applique dans les deux cas.

On peut inclure des paramètres de régression dans l'équation (4.1) en remplaçant y_i par l'écart entre y_i et la droite de régression. Tam (1987) a élargi ce concept en considérant un modèle où le processus stochastique correspondant est déterminé par un modèle d'espace d'états des coefficients de régression en évolution.

Pour maximiser la fonction de vraisemblance (4.1) par rapport aux paramètres inconnus, il faut une méthode itérative. Nous sautons ici les détails de la méthode utilisée dans l'exemple de la Section 5 puisqu'on est encore à élaborer des méthodes efficaces.

Une fois qu'on a estimé les paramètres, on peut calculer des valeurs lissées pour le vecteur d'états, $\hat{z}_{i|T} = E(z_i|X_T)$ pour $T > i$, à l'aide des formules d'extrapolation rétrospective définies par le filtre de Kalman; voir Harvey (1984). Par exemple, si $y_i = \theta + e_i$ comme en (3.1), l'extrapolation rétrospective nous permettra d'écrire $y_i = \hat{\theta}_{i|T} + \tilde{e}_{i|T}$, de sorte que $\tilde{\theta}_{i|T}$ devient l'estimation lissée de la moyenne au temps i une fois X_T connu.

Pour calculer l'erreur type de l'estimation lissée, il faut se rappeler, à plus forte raison lorsqu'on calcule l'erreur type de la moyenne au temps i une fois X_T connu.

L'estimation lissée par des simulations de Monte Carlo. Il produit un ensemble de variables aléatoires normales multidimensionnelles dont la moyenne est déterminée par l'estimation la plus vraisemblable des paramètres et la variance déterminée par l'inverse de la matrice d'information de Fisher estimée. Il estime ensuite $E(P_{i|T})$ et $\text{Var}(\hat{z}_{i|T})$, où l'espérance mathématique et la variance sont calculées pour l'ensemble des valeurs de paramètres estimées. La somme de ces deux éléments est la matrice des covariances estimées du vecteur d'états estimé. Cette méthode suppose un grand échantillon, de sorte qu'on peut dire que la distribution d'échantillonnage des estimations des paramètres est approximativement normale.

Dans l'exemple de la Section 5, nous calculons une valeur approchée pour l'écart type des erreurs d'échantillonnage des estimations lissées en ne tenant pas compte de la variation causée par l'estimation de certains paramètres du modèle. Nous comparons ensuite ces valeurs aux erreurs types de la distribution d'échantillonnage établie à l'aide de données simulées.

5. ANALYSE DE DONNÉES

Dans cette section, nous allons voir l'incidence des erreurs de sondage sur les valeurs estimées des paramètres d'un modèle autorégressif du premier degré comprenant des termes de régression. Dans l'exemple que nous présentons, nous supposons que les erreurs de sondage sont indépendantes d'une enquête à l'autre. Le cadre d'analyse que nous venons d'exposer permet de traiter les cas plus complexes où il y a corrélation des erreurs de sondage et où la caractéristique de population est expliquée par des modèles ARMA de degré supérieur. Nous avons choisi cet exemple pour montrer que le fait de tenir compte des erreurs de sondage peut avoir des conséquences appréciables même pour un modèle relativement simple.

Pour notre exemple, nous nous sommes servi des données fournies par des résidents de la Saskatchewan à l'enquête sur les voyages des Canadiens (EVC). L'EVC est réalisée par Statistique Canada dans le but de recueillir des statistiques descriptives sur les habitudes de voyage des résidents canadiens et les caractéristiques des voyageurs canadiens. Elle est un sous-échantillon constant où il y a six groupes de renouvellement. Or, l'EVC revient au plus quatre fois par année et utilise au moins un groupe de renouvellement mais peut en utiliser jusqu'à trois. Les groupes utilisés pour les trimestres où il y a une EVC sont choisis de manière qu'il n'y ait pas de chevauchement d'une fois à l'autre.

Les erreurs de sondage sont supposées indépendantes. Cette hypothèse n'est pas parfaitement exacte. En effet, l'EPA est une enquête à plusieurs degrés et les unités primaires d'échantillonnage (UPÉ) ne se succèdent pas aussi rapidement que les groupes de renouvellement dans l'échantillon. Les mêmes UPÉ servent à plusieurs reprises. Ainsi, bien que l'échantillon de

On peut aussi utiliser les modèles d'états habituels lorsque les erreurs de mesure touchant les données d'entrée sont indépendantes. C'est ce que nous faisons dans l'exemple de la Section 5, où nous montrons ce qui arrive aux estimations de paramètres lorsque les erreurs de sondage sont prises en considération.

L'estimation de ces paramètres par la méthode du maximum de vraisemblance lorsqu'il y a corrélation entre les erreurs de sondage est un sujet qui n'a jamais été approfondi. Dans le cas d'un modèle avec des observations stationnaires unidimensionnelles $\{y_t\}$ Scott, Smith et Jones (1977) ont proposé d'utiliser la fonction d'autocovariance estimée des observations, $\{y_t\}$ pour estimer les paramètres du processus ARMA. En l'occurrence, le modèle de données s'écrit $y_t = \theta_t + e_t$. On peut estimer les variances et les covariances des erreurs de sondage, $\{e_t\}$ à l'aide de méthodes fondées sur le plan; voir par exemple Wolter (1985). Comme il n'existe pas vraiment d'ouvrage où l'on parle de l'estimation efficace des autocovariances des erreurs de sondage dans l'hypothèse d'une série stationnaire, on recourra dans la pratique à des méthodes improvisées. Il serait donc utile de pousser la recherche sur la modélisation des erreurs de sondage. Dans l'exemple de la Section 5, cette lacune ne pose pas de problème car nous avons pu supposer que les erreurs de sondage étaient indépendantes.

En supposant que nous connaissons l'autocovariance de $\{e_t\}$ nous pouvons estimer l'autocovariance de $\{\theta_t\}$ par la formule $\text{Cov}(\theta_t, \theta_{t-s}) = \text{Cov}(y_t, y_{t-s}) - \text{Cov}(e_t, e_{t-s})$. Cependant, cette méthode n'est pas parfaitement efficace (Smith, 1978). En outre, elle ne tiendrait pas compte des erreurs de sondage non stationnaires.

Miazaki (1985) a examiné le cas où $\{\theta_t\}$ suit un processus ARMA $(p, 0)$. Elle a aussi supposé que $\{e_t\}$ suit un processus ARMA $(0, q)$ qui peut être estimé directement à partir de l'enquête. Ensuite, elle a exprimé les observations $\{y_t\}$ suivant un processus ARMA $(p, p+q)$ qu'elle a estimé par des méthodes du maximum de vraisemblance assujéties à des contraintes. Dans les modèles d'espace d'états, on peut parfois représenter la non-stationnarité des erreurs de sondage en substituant des matrices non homogènes à V_t , l'inqui est la matrice des variances des «perturbations» aléatoires découlant de l'équation de transition (3.6b). Dans l'équation (3.7) par exemple, on substituerait s^2 à s_t^2 pour tenir compte des erreurs de sondage non homogènes. C'est ce que nous faisons dans l'exemple de la Section 5.

De façon générale, Harvey et Phillips (1979) écrivent comme suit la fonction exacte de vraisemblance pour les modèles d'états définis par les équations (3.5). Soit

$$y_{t|t-1} = E(y_t|Y_{t-1}) = H_t' z_{t|t-1}$$

et

$$R_t = \text{Var}(y_t|Y_{t-1}) = H_t' P_{t|t-1} H_t + U_t,$$

la fonction de vraisemblance logarithmique pour $Y_T^T = (y_1^T, \dots, y_T^T)$ est

$$\log f(Y_T) = (1/2) \sum_{t=1}^T \log |R_t| - (1/2) \sum_{t=1}^T (y_t - y_{t|t-1})' R_t^{-1} (y_t - y_{t|t-1}). \quad (4.1)$$

Les paramètres inconnus de l'équation (4.1) sont compris dans $y_{t|t-1}$ et R_t . Suivant l'algorithme utilisé pour maximiser (4.1) par rapport aux paramètres inconnus, on pourrait devoir calculer les dérivées première et seconde de (4.1) par rapport à ces paramètres. Pour cela, on doit normalement calculer les dérivées de $z_{t|t-1}$ et de $P_{t|t-1}$. Celles-ci peuvent être calculées à l'aide des formules récursives définies en (3.8) et en (3.9). Par exemple, (3.9c) donne $\partial z_{t|t-1} = (\partial F_t) z_{t-1|t-1} + F_t (\partial z_{t-1|t-1})$. On peut déterminer de la même façon les autres expressions tirées de (3.8) et (3.9).

Les nouvelles valeurs de la moyenne et de la variance pour le vecteur d'états au temps t une fois que les observations au temps t sont connues sont déterminées par les équations suivantes:

$$E(z_t|Y_t) = \hat{z}_{t|t}$$

$$= \hat{z}_{t|t-1} + P_{t|t-1}H_t'(H_t'P_{t|t-1}H_t + U_t)^{-1}(y_t - H_t'\hat{z}_{t|t-1})$$

(3.10a)

$$\text{Var}(z_t|Y_t) = P_{t|t} = P_{t|t-1} - P_{t|t-1}H_t'(H_t'P_{t|t-1}H_t + U_t)^{-1}H_t'P_{t|t-1}$$

(3.10b)

Les équations (3.9) et (3.10) sont les équations bien connues du filtre de Kalman. La formulation utilisée ici est essentiellement bayésienne mais on peut obtenir des résultats équivalents au moyen de projections orthogonales; voir Young (1984).

Nous pouvons constater à quel point les équations du filtre de Kalman simplifient les calculs dans les enquêtes par sondage en comparant les équations (3.9) et (3.10) avec celle obtenue par R.G. Jones (1980) (équation 3.5). Notons que pour en arriver à cette équation, Jones a dû inverser une matrice dont les dimensions étaient déterminées par le vecteur des estimations d'enquête.

Le filtre de Kalman peut aussi servir à calculer des estimations lissées que l'on définit par $E(z_t|Y_t)$ pour $T > t$. Pour plus de détails sur cette extrapolation rétrospective, voir Harvey (1984).

Remarques

1. Bien que le filtre de Kalman suppose un modèle de population infinie, il arrive souvent que, grâce au théorème de la limite centrée, les erreurs de sondage suivent une distribution approximativement normale lorsque l'enquête repose sur un grand échantillon. De la même façon, comme les estimateurs lissés de $\{\theta_t'\}$ sont identiques à ceux calculés par R.G. Jones (1980) en (3.5a), il s'agit là des estimateurs linéaires à EQM minimum même en l'absence des hypothèses de normalité.

2. Le modèle d'espace d'états peut tenir compte des observations manquantes à une période donnée. Si on ne connaît pas y_t au temps t , les équations de mise à jour analogues à (3.9) deviennent $\hat{z}_{t|t} = \hat{z}_{t|t-1}$ et $P_{t|t} = P_{t|t-1}$ comme dans R. H. Jones (1980). Cependant, les estimations lissées correspondant aux observations manquantes dépendront largement du modèle choisi étant donné l'absence d'estimations d'enquête. Il y a donc ici une forte probabilité de mal définir le modèle.

3. La fonction de vraisemblance, dont nous servons dans la Section 4 pour calculer les estimations les plus vraisemblables des paramètres inconnus, peut être définie en situation de non-réponse à l'aide de la méthode proposée par R. H. Jones (1980). Toutefois, les données manquantes tendront à accroître les erreurs types des valeurs estimées des paramètres. Dans l'exemple de la Section 5, il y a des observations manquantes à une période donnée.

4. ESTIMATION DES PARAMÈTRES DANS UN MODÈLE D'ESPACE D'ÉTATS

Lorsque les données proviennent du modèle ARMA (équation 3.7) et que les paramètres α , β , et σ^2 sont inconnus, on peut calculer l'estimation la plus vraisemblable des paramètres inconnus à l'aide de la fonction de vraisemblance issue du modèle d'espace d'états. C'est la méthode au 'ont proposée Harvey et Phillips (1979). R. H. Jones (1980) et d'autres.

dans le cas d'une enquête où le plan d'échantillonnage correspond à l'échantillonnage avec renouvellement à deux degrés d'Eckler, où l'estimation d'enquête pour θ_t est représentée par \bar{y}_t , qui est la moyenne calculée pour l'ensemble des personnes qui participent à l'enquête pour la t -ième période.

Nous pouvons exprimer cela sous forme de modèle d'espace d'états en posant

$$(3.8) \quad F_t = \begin{bmatrix} \alpha_1 & 1 & 0 & 0 & 0 \\ \alpha_2 & 0 & 1 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad G_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\beta^* \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad V_t = \begin{bmatrix} \sigma^2 & 0 \\ 0 & s_2 \end{bmatrix},$$

$U_t = 0$ et $H_t' = (1 \ 0 \ 0 \ 1 \ 0)$. Les trois premiers éléments du vecteur d'états se rapportent au processus $\{\theta_t\}$ tandis que les deux derniers se rapportent au processus $\{e_t\}$.

Notons que les modèles d'espace d'états tiennent compte de l'erreur de mesure, qui est représentée par \bar{w}_t dans l'équation (3.6a). Or, à moins qu'il s'agisse d'une enquête sans participation répétée avec erreurs d'échantillonnage indépendantes, on ne peut se servir des termes d'erreur de mesure pour modéliser l'erreur de sondage. C'est pourquoi nous avons intégré l'erreur de mesure (de sondage) dans le vecteur d'états.

À partir du modèle général d'espace d'états, il est possible de déterminer les équations du filtre de Kalman. Si, comme dans Meinhold et Singpurwalla (1983), nous définissons la distribution conditionnelle de z_{t-1} étant donné X_{t-1} comme $N(\bar{z}_{t-1|t-1}, P_{t-1|t-1})$, il est possible de construire des relations récursives pour $\bar{z}_{t|t}$ et $P_{t|t}$ Harvey (1984) montre que ces relations sont l'équivalent du filtre de Kalman.

De façon générale, le filtre de Kalman se compose de deux parties. La première est une prévision une étape à l'avance du vecteur d'états et de sa covariance; la seconde est une mise à jour de la matrice des moyennes et des covariances du vecteur d'espace d'états une fois que les nouvelles observations sont connues.

Utilisant la même notation qu'en (3.6), nous posons $X_t = y_t$ et $X_{t+1}' = (X_t', y_{t+1}')$, alors la moyenne et la variance de la prévision une étape à l'avance sont

$$(3.9a) \quad E(z_1) = \bar{z}_{1|0}$$

$$(3.9b) \quad \text{Var}(z_1) = P_{1|0}$$

$$(3.9c) \quad E(z_t|X_{t-1}) = \bar{z}_{t|t-1} = F_t' \bar{z}_{t-1|t-1}$$

$$(3.9d) \quad \text{Var}(z_t|X_{t-1}) = P_{t|t-1} = F_t' P_{t-1|t-1} F_t' + G_t' V_t G_t'$$

Harvey et Phillips (1979) ont décrit une méthode permettant d'exprimer le modèle ARMA (p,q), défini par:

(3.7)
$$y_t - \alpha_1 y_{t-1} - \dots - \alpha_p y_{t-p} = \epsilon_t - \beta_1 \epsilon_{t-1} - \dots - \beta_q \epsilon_{t-q},$$

où les ϵ_t 's sont indépendants et distribués suivant une loi $N(0,\sigma^2)$, sous forme de modèle d'espace d'états. La dimension de z_t est $r = \text{MAX}(p,q+1)$. Lorsque cela est nécessaire, on ajoute des zéros à $\bar{\alpha} = (\alpha_1, \dots, \alpha_p)$ ou à $\bar{\beta} = (\beta_1, \dots, \beta_q)$ pour obtenir des vecteurs de dimension r . La matrice U_t est définie comme une matrice nulle. Le modèle ARMA(p,q) équivaut à (3.6) lorsque $H_t' = (1, 0, \dots, 0), G_t' = (1, -\beta_1, \dots, -\beta_{r-1})$ et

$$F_t = \left[\begin{array}{c|c} \alpha_r & \alpha_r \\ \hline \alpha_{r-1} & \alpha_{r-1} \\ \vdots & \vdots \\ \alpha_1 & \alpha_1 \\ \hline I_{r-1} & O' \end{array} \right],$$

où I_{r-1} est la matrice unité de dimensions $(r-1) \times (r-1)$ O' est un vecteur ligne formé de zéros. Selon cette méthode, le vecteur d'états $z_t = (z_{1t}, \dots, z_{rt})'$ est défini comme suit:

$$z_{1t} = \alpha_1 y_{t-1} + \alpha_{t+1} y_{t-2} + \dots + \alpha_r y_{t-(r-t+1)} \\ - \beta_{t-1} \epsilon_t - \beta_t \epsilon_{t-1} - \beta_{t+1} \epsilon_{t-2} - \dots - \beta_r \epsilon_{t-(r-t)},$$

pour $t = 2, 3, \dots, r$ et $z_{1t} = y_t$ tel que défini en (3.7). Une condition nécessaire pour la stationnarité est que $\text{Var}(z_t) = \text{Var}(z_{t-1})$ pour tous les t . D'après l'expression (3.6b), nous voyons que cette condition implique:

$$\text{Var}(z_t) = F' \text{Var}(z_t) F + G G',$$

où $V_t \equiv V$ est constante pour tous les t . Pearlman (1980) a indiqué que cela pouvait servir à définir les conditions initiales pour z_1 .

On peut souvent inclure le processus des erreurs de sondage dans le modèle d'espace d'états lorsqu'il est possible de supposer une structure pour ces erreurs de sondage. Nous l'avons déjà démontré en ce qui concerne le modèle de Blight et Scott (1973). Scott et Smith (1974) et Mizaki (1985) ont examiné une série de modèles qui étaient des cas particuliers d'un processus (θ_t) ARMA(p,q), pour $\{\epsilon_t\}$ et d'un processus ARMA(p^*,q^*) et ont étudié les observations scalaires qui satisfont $y_t = \theta_t + \epsilon_t$. Les modèles d'espace d'états pour ces processus peuvent être formulés suivant la méthode de Harvey et Phillips décrite ci-dessus; en l'occurrence, le vecteur d'états z_t est le résultat de l'enchaînement des vecteurs d'états de chaque processus ARMA. Supposons, par exemple, que $\{\theta_t\}$ suit un processus ARMA(3,0) avec des paramètres ($\alpha_1, \alpha_2, \alpha_3$) et une variance de modèle σ^2 et que $\{\epsilon_t\}$ suit un processus ARMA(0,1) avec un paramètre β^* et une variance de modèle s^2 . La seconde partie de cette hypothèse est plausible

Dans le modèle d'espace d'états, deux processus se déroulent simultanément. Le premier, qui est le système d'observation, décrit en détail comment les observations dépendent de l'état des paramètres du processus dans la période observée. Le second, qui est le système de transition, décrit en détail l'évolution des paramètres. On peut formuler les modèles d'espace d'états de la façon suivante. L'équation d'observation s'écrit:

$$y_t = H_t z_t + \bar{w}_t, \tag{3.6a}$$

et l'équation de transition s'écrit:

$$z_t = F_{t-1} z_t + G_t \epsilon_t, \tag{3.6b}$$

où z_t est un vecteur d'états ($r \times 1$), H_t est une matrice fixe ($n_t \times r$), F_t est une matrice de transition fixe ($r \times r$), G_t est une matrice fixe ($r \times m$) et \bar{w}_t et ϵ_t sont des bruits aléatoires indépendants de moyenne nulle et de covariances $E(\bar{w}_t \bar{w}_t') = U_t$ et $E(\epsilon_t \epsilon_t') = V_t$.

À titre d'exemple, nous allons transformer le modèle analysé par Blight et Scott (1973) en un modèle d'espace d'états. Blight et Scott ont considéré les données du plan d'échantillon-nage avec renouvellement à un degré de Patterson (1950). Ils ont défini y_t' comme la moyenne des nouvelles unités au temps t , et x_t' et y_t' comme les moyennes, au temps t et $t - 1$, respectivement, des unités ayant déjà participé. Ils ont supposé que y_t'' et y_t' — $\rho x_t'$ sont des observations indépendantes au temps t , où ρ est la corrélation entre les réponses fournies par la même personne d'une fois à l'autre. Ils ont aussi supposé que la moyenne $\{\theta_t\}$ suit un processus autorégressif du premier degré.

Soit le vecteur d'états $z_t' = (\theta_t, \theta_{t-1})$. Nous pouvons écrire l'équation d'observation comme suit:

$$\begin{bmatrix} y_t'' \\ y_t' - \rho x_t' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\rho \end{bmatrix} \begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} + \begin{bmatrix} \omega_{1t} \\ \omega_{2t} \end{bmatrix},$$

où $(\omega_{1t}, \omega_{2t})'$ a une matrice des covariances diagonale.

L'équation de transition s'écrit:

$$\begin{bmatrix} \theta_t \\ \theta_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_{t-1} \\ \theta_{t-2} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \epsilon_t,$$

où ϵ_t est distribué suivant une loi $N(0, \sigma^2)$. Par conséquent, nous pouvons exprimer le modèle de Blight-Scott sous forme de modèle d'espace d'états.

l'échantillon et la variance totale. Lorsque l'erreur de sondage diminue, l'importance de Θ_{t-1} diminue aussi et l'estimation de θ_t dans (3.3a) se compose principalement de y_t , qui est l'estimation établie à l'aide des données d'enquête. Par conséquent, l'estimateur de θ_t est convergent selon le plan lorsque y_t l'est.

En revanche, lorsque l'erreur de sondage augmente, l'estimation de θ_t est déterminée principalement par la prévision linéaire de Θ_{t-1} . L'efficacité relative de l'estimateur, θ_t , dans (3.3a) est définie par $1/(1 - \pi_t)$, où π_t est défini en (3.3c). On obtient les gains d'efficacité les plus élevés lorsque l'erreur de sondage est grande par rapport à $\sigma_{\theta_t}^2$, qui est la variance des perturbations du processus.

Scott et Smith (1974) et R.G. Jones (1980) se sont aussi intéressés aux enquêtes avec participation répétée. Jones a formulé son analyse dans les termes suivants. Soit Θ_t une variable aléatoire normale multidimensionnelle avec une moyenne nulle et une matrice des covariances U_t . Par ailleurs, on peut généraliser les observations au temps t en un vecteur d'estimations élémentaires, y_t . La distribution conditionnelle de $Y_t = (y_t', \dots, y_t')$ étant donné Θ_t est supposée être de la forme:

$$Y_t = X_t' \Theta_t + e_t, \tag{3.4}$$

où X_t est une matrice fixe formée de zéros et de uns qui fait le lien entre les paramètres et les observations et e_t est l'erreur de sondage, qui est supposée être une variable aléatoire normale multidimensionnelle avec une moyenne nulle et une matrice des covariances U_t .

À l'aide d'arguments conditionnels, on peut montrer que la meilleure estimation de θ_t étant donné Y_t est:

$$E(\theta_t | Y_t) = \Theta_t = (X_t' U_t^{-1} X_t + V_t^{*-1})^{-1} X_t' U_t^{-1} Y_t \tag{3.5a}$$

avec comme variance

$$\text{Var}(\theta_t | Y_t) = (X_t' U_t^{-1} X_t + V_t^{*-1})^{-1}. \tag{3.5b}$$

Ce résultat est très général. Si jamais le modèle stochastique pour Θ_t était très faible, l'inverse de V_t^* serait approximativement 0, ce qui donnerait l'ELNVM défini en (2.2a). R.G. Jones (1980) a obtenu les équations (3.5) à l'aide des moindres carrés stochastiques, de sorte que l'estimateur Θ_t est l'estimateur linéaire à erreur quadratique moyenne minimum (EQMM) même en l'absence des hypothèses de normalité.

Si on appliquait directement (3.5), il faudrait inverser des matrices qui sont de même dimension que le vecteur de toutes les estimations élémentaires pour toutes les périodes. Calculer l'inverse de telles matrices pourrait donner des résultats incertains. Cependant, on peut souvent redéfinir l'expression (3.5) à l'aide des modèles d'états, qui permettent de décrire de nombreux modèles de séries chronologiques. Voir Harvey (1984) pour une analyse de ces modèles. Comme nous allons le voir ci-dessous, l'utilisation des modèles d'états nous évite d'inverser de grandes matrices. Pour tirer parti de la réduction de dimensions que permet l'utilisation des modèles d'états, il est nécessaire de définir une structure pour $\{\theta_t\}$ et $\{e_t\}$. On pourrait par exemple choisir un processus autorégressif de moyennes mobiles (ARMA) qui n'est pas nécessairement homogène dans le temps; ce genre de structure est souvent utilisé dans les applications de séries chronologiques.

Pour des applications comme le calcul d'estimations pour petites régions, où les tailles d'échantillons sont relativement faibles, il peut être utile de modéliser les variances de l'erreur de sondage, U_t à l'aide de processus ARMA. On ne procède pas de cette façon habituellement dans le cas des enquêtes répétées. L'utilisation de modèles ARMA faciliterait aussi l'application directe de l'équation (3.5) lorsque les dimensions de, V_t^* et U_t sont grandes et que les inverses sont numériquement instables.

Dans la présente section, nous allons montrer comment les hypothèses du modèle stochastique peuvent aussi permettre d'obtenir des estimateurs fondés sur un modèle et convergents selon un plan. Dans la Section 4, il est question de l'estimation des paramètres du modèle par la méthode du maximum de vraisemblance. Comme une définition erronée du modèle peut entraîner des biais appréciables, on devrait recourir à des méthodes de vérification d'hypothèses pour s'assurer que le modèle est conforme aux données. Le modèle devrait en outre refléter une connaissance approfondie du phénomène en question.

Voyons tout d'abord le cas où les erreurs de sondage sont indépendantes. (Cette hypothèse est vraisemblable dans le cas des enquêtes sans participation répétée avec de faibles taux de sondage.) En l'occurrence, l'ELNVM de θ_i est $\theta_i = y_i$. Or, en appliquant un modèle stochastique à la suite des paramètres, $\{\theta_i\}$, il est possible de réduire l'erreur quadratique moyenne de l'estimateur.

Scott et Smith (1974) ont proposé le modèle suivant pour les enquêtes sans participation répétée. Leur modèle des estimations d'enquête au temps t s'écrit comme suit:

$$y_t = \theta_t + e_t \tag{3.1}$$

où les résidus e_t sont indépendants et distribués suivant une loi normale de moyenne nulle et de variance indépendante $N(0, S_t^2)$. Ils ont supposé que la suite des paramètres, $\{\theta_t\}$ peut être modélisée sous la forme suivante (à la condition que $\Theta_{t-1} = (\theta_1, \dots, \theta_{t-1})$,

$$\theta_t = \bar{\alpha}'_t \Theta_{t-1} + e_t, \tag{3.2}$$

où les e_t 's sont indépendants et distribués suivant une loi normale de moyenne nulle et de variance $N(0, S_t^2)$ et sont indépendants de $\{e_t\}$, et où $\bar{\alpha}_t$ est un vecteur de constantes de dimension $(t-1)$.

En règle générale, au temps $t-1$, nous avons $Y'_{t-1} = (y_1, \dots, y_{t-1})$, à la condition $\Theta_{t-1} \sim N(\bar{\Theta}_{t-1}, V'_{t-1})$. Par des arguments conditionnels, nous pouvons montrer que

$$E(\theta_t | y_t) = \theta_t = \pi_t(\bar{\alpha}'_t \Theta_{t-1}) + (1 - \pi_t)y_t \tag{3.3a}$$

et que

$$\text{Var}(\theta_t | y_t) = (1 - \pi_t)S_t^2, \tag{3.3b}$$

où

$$\pi_t = \frac{\text{Var}(y_t | \theta_t)}{\text{Var}(y_t)} = \frac{S_t^2}{S_t^2 + \bar{\alpha}'_t V'_{t-1} \bar{\alpha}_t + \sigma_t^2 + S_t^2}. \tag{3.3c}$$

Notons que l'estimateur défini en (3.3a) est la moyenne pondérée de deux éléments: le premier représente la meilleure prévision linéaire de θ_t étant donné la valeur précédente de $\bar{\Theta}_{t-1}$; et le second représente la meilleure estimation de θ_t établie à l'aide des données de l'enquête. L'importance de chaque terme est déterminée par π_t , qui est le rapport entre la variance de

différentes peut se décomposer en deux éléments: i) la corrélation entre les unités du second degré (USE) tirées des unités primaires d'échantillonnage (UPF) et ii) la corrélation entre les moyennes d'UPF de périodes successives. Si on suppose que ces deux formes de corrélation suivent un processus autorégressif du premier degré, la forme de l'ELNVM correspond à la forme générale définie par Patterson (1950).

Tikkiwal (1979) et d'autres ont examiné ce qui arriverait si on assumait l'hypothèse ci-dessus. Tikkiwal a conclu que si on supposait l'existence d'une structure de corrélation entièrement générale, la forme simple de l'ELNVM disparaîtrait et il faudrait recourir, en pratique, à l'approximation. Rao et Graham (1964) et Gurney et Daly (1965) ont proposé l'utilisation d'estimateurs composites, qui sont une approximation des estimateurs optimaux. Ces estimateurs s'utilisent facilement et ont une efficacité relative élevée. Pour une étude de l'utilisation de ces estimateurs, voir Binder et Hidiroglou (1988). Gurney et Daly (1965) ont de plus étendu les résultats de Patterson (1950) à un modèle linéaire. Ils ont défini la notion d'«estimation élémentaire». Il s'agit d'une estimation fondée sur des données se rattachant à une période précise et recueillies auprès d'un groupe de personnes qui sont retranchées de l'échantillon ou y sont intégrées ensemble. On peut exprimer l'espérance mathématique de ces estimations élémentaires comme une combinaison linéaire des paramètres de population, $\{\theta_i\}$. Lorsqu'on connaît la structure de corrélation, on peut calculer l'ELNVM à l'aide de la théorie générale des modèles linéaires. Afin de mathématiser cette analyse, posons y_{ij} comme la j -ème estimation élémentaire rattachée à la période $E(y_{ij}) = \theta_i$. Si Y et Θ sont des vecteurs ayant pour éléments y_{ij} et θ_i respectivement, nous pouvons écrire :

$$Y = X'\Theta + e, \tag{2.1}$$

où X est une matrice fixe ($n \times T$) formée de zéros et de uns, $E(e) = 0$ et $E(ee') = U$, qui est la matrice connue des variances-covariances des estimations élémentaires. Donc, l'ELNVM est défini par l'équation suivante:

$$\Theta = (X'U^{-1}X)^{-1}X'U^{-1}Y, \tag{2.2a}$$

où

$$\text{Var}(\Theta) = (X'U^{-1}X)^{-1}. \tag{2.2b}$$

Ces résultats supposent qu'à chaque nouvelle enquête il faut mettre à jour toutes les estimations précédentes. Toutefois, comme les estimations tirées des enquêtes plus reculées ont souvent un effet beaucoup moindre que les estimations des enquêtes récentes, les estimateurs composites, comme ceux proposés par Gurney et Daly (1965), sont plus faciles à utiliser et ont une efficacité relative élevée. Binder et Hidiroglou (1988) ont étudié la pertinence de ces méthodes ainsi que leur application dans un certain nombre d'enquêtes. De façon générale, ils ont constaté que l'on peut obtenir de bons résultats avec des estimateurs composites, pourvu que les biais de renouvellement ne soient pas trop élevés.

3. DISSOCIATION DU SIGNAL ET DU BRUIT

Les économistes et les sociologues ont souvent tendance à considérer les paramètres $\{\theta_i\}$ comme des variables aléatoires dans leurs modèles stochastiques (Smith 1978). Cependant, si on ne tient pas compte des erreurs d'échantillonnage liées aux données d'entrée, les valeurs estimées des paramètres du modèle stochastique se trouvent biaisées.

Lorsqu'on suppose un modèle ARMA en présence d'erreurs de sondage, on peut recourir aux modèles d'espace d'états pour calculer les estimations les plus vraisemblables des paramètres inconnus. Notons que cette méthode peut être assimilée à une méthode empirique de Bayes. Nous supposons que les erreurs de sondage peuvent être décrites par un processus ARMA jusqu'à un facteur multiplicatif. C'est ce que nous voyons dans la Section 4.

Dans la Section 5, nous donnons un exemple de ce modèle en nous servant des données de l'enquête sur les voyages des Canadiens. Cet exemple montre comment le fait de tenir compte des erreurs de sondage influe sur les valeurs estimées des paramètres du modèle. Par la même occasion, nous calculons des estimations lissées suivant les hypothèses du modèle. Comme, dans cet exemple, les erreurs de sondage sont indépendantes, nous n'avons pas à exposer tout l'appareil mathématique de la formulation générale. Cependant, l'exemple montre que le fait de ne pas tenir compte des erreurs de sondage peut avoir des conséquences appréciables même dans ces conditions.

La Section 6 renferme les conclusions de l'étude.

2. ESTIMATION LINÉAIRE NON BIAISÉE À VARIANCE MINIMUM DANS LES ENQUÊTES RÉPÉTÉES AVEC PARTICIPATION RÉPÉTÉE DE CERTAINES UNITÉS

Dans cette section, nous faisons un survol des articles où l'on considère la valeur d'une caractéristique de population comme la moyenne ou le total comme une constante inconnue. Dans la Section 3, nous étudions le cas où l'on suppose un modèle stochastique pour la caractéristique de population.

Dans les enquêtes à participation répétée, où des personnes sont appelées à participer plus d'une fois à la même enquête, les erreurs d'échantillonnage sont habituellement corrélées d'une fois à l'autre. Il peut aussi y avoir corrélation dans les enquêtes à plusieurs degrés où quelques-unes des unités d'échantillonnage du premier degré participent plus d'une fois, même si en définitive il ne s'agit pas des mêmes répondants.

Les estimateurs qui ne tiennent pas compte de ces corrélations et ne reposent que sur les données recueillies dans une seule période de référence sont en règle générale inefficaces comparativement à l'estimateur linéaire non biaisé à variance minimum (ELNVM). L'efficacité relative dépend du degré de corrélation entre les erreurs d'échantillonnage d'une enquête à l'autre. Lorsque le degré de corrélation est nul, comme c'est le cas dans l'exemple de la Section 5, l'ELNVM est simplement l'estimateur fondé sur les données d'une seule période de référence. C'est Jessen (1942) qui, le premier, a élaboré une théorie générale pour les enquêtes répétées auxquelles certaines unités participent plus d'une fois. Il a examiné en détail le cas particulier d'un échantillon aléatoire simple tiré d'une population infinie, où la corrélation pour les personnes décroît exponentiellement d'une période à l'autre. À chaque période d'enquête, un certain nombre de personnes sont retranchées de l'échantillon de la période précédente puis d'autres y sont ajoutées. Les données recueillies concernent uniquement la période courante. Patterson a calculé l'ELNVM pour ce cas particulier.

Les hypothèses fondamentales de Patterson (1950) ont été reprises puis élargies. Eckler (1955) a désigné le plan de Patterson comme un échantillonnage avec renouvellement à un degré. Il a déterminé l'ELNVM pour le cas où des personnes fournissent des renseignements pour deux périodes successives, ce qu'il a appelé l'échantillonnage avec renouvellement à deux degrés. Il a aussi calculé l'ELNVM pour des plans d'échantillonnage avec renouvellement de degré supérieur.

Rao et Graham (1964) ont assoupli l'hypothèse de la population infinie en intégrant le facteur de correction pour population finie dans la variance de l'erreur de sondage. Singh (1968) a été le premier à considérer des plans à plusieurs degrés. Il a examiné des plans de sondage à deux degrés en supposant que la corrélation entre les réponses fournies à des périodes

Enquêtes répétées – Modélisation et estimation

D.A. BINDER et J.P. DICK¹

RÉSUMÉ

Les auteurs examinent brièvement l'estimation de la moyenne d'une caractéristique pour une population à différentes périodes à partir d'une série d'enquêtes successives. En définissant un modèle paramétrique stochastique pour ces moyennes, il est possible d'estimer les paramètres et d'obtenir des estimateurs des moyennes proprement dits. Les auteurs exposent le cas où les moyennes de population suivent un processus autorégressif de moyennes mobiles (ARMA) et où les erreurs de sondage peuvent aussi être exprimées par un tel processus. Enfin, les auteurs ont recours à un exemple où ils utilisent des données de l'enquête sur les voyages des Canadiens.

MOTS CLÉS: Filtre de Kalman; enquêtes avec participation répétée; modèles d'états; modèles de séries chronologiques; estimations pour petites régions.

1. INTRODUCTION

Une enquête qui revient périodiquement permet l'application de méthodes d'estimation et d'analyse qui ne conviennent pas pour une enquête unique. Par exemple, l'efficacité des méthodes d'estimation utilisées à une occasion peut dépendre des données recueillies les fois précédentes. Cela se produit lorsque des unités d'échantillonnage participent à deux enquêtes successives et que, par conséquent, on peut établir une corrélation entre les erreurs de sondage d'une période à l'autre. De même, les utilisateurs de données définissent souvent des modèles pour les séries d'estimations tirées d'enquêtes répétées. Par exemple, on suppose souvent un modèle autorégressif de moyennes mobiles (ARMA). Cependant, la plupart des méthodes qui permettent actuellement d'estimer les paramètres de ce modèle supposent que les données d'entrée ne sont pas exposées à l'erreur de sondage.

Dans cet article, nous élaborons des méthodes permettant d'estimer ces paramètres lorsque que les données renferment des erreurs de sondage. La structure de covariance des erreurs que nous considérons comprend des cas où les erreurs de sondage sont corrélées dans le temps. Lorsque l'on suppose un modèle de ce genre pour décrire le mouvement des caractéristiques de la population, il est possible de déterminer l'estimateur linéaire à erreur quadratique moyenne minimum (EQMM). Cet estimateur admet la structure de modèle que l'estimateur linéaire non biaisé à variance minimum (ELNVM) n'admet pas. Nous reviendrons sur l'ELNVM dans la Section 2.

Blight et Scott (1973), Scott et Smith (1974), Scott, Smith et Jones (1977), R.G. Jones (1980) et d'autres encore ont examiné les conséquences de modèles stochastiques de ce genre pour les moyennes de population dans le temps. Les résultats de leur analyse ainsi qu'une formulation plus générale faisant intervenir des modèles d'espace d'états et des filtres de Kalman sont étudiés dans la Section 3 pour le cas où le modèle stochastique pour les caractéristiques de population est défini entièrement. Ces méthodes peuvent être élaborées dans le cadre d'un modèle de type bayésien, où la distribution a priori est définie entièrement.

¹ D.A. Binder et J.P. Dick, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

- ALLEN, R., CLAMPET, G., DUNKERLEY, C., TORTORA, R., et VOGEL, F. (1983). Framework for the Future. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- BYNUM, H., DOWDY, W., HANUSCHAK, G., HUDSON, C., MURPHY, R., STEINBERG, J., et VOGEL, F.A. (1985). Crop Reporting Board Standards. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- COCHRAN, W.G (1977). *Sampling Techniques*. New York: Wiley.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., et PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.
- HOUSEMAN, E.E. (1975). Area Frame Sampling in Agriculture. SRS Report No. 20. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- KUO, L. (1986). Composite Estimation of Totals for Livestock Surveys. SF & SRB Staff Report No. 92. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. SF & SRB Staff Report No. 80. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., et FULLER, W.A. (1988). Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes. *Techniques d'enquête*, 14, 63-73.

BIBLIOGRAPHIE

Tableau 6

Matrices des covariances estimées pour les estimateurs à base aréolaire à segment pondéré, fondées sur des données des groupes de renouvellement

	1982	1983	1984	1985	1986
Porcs					
1982	0.899	0.436	0.283	0.180	0.124
1983	0.436	0.857	0.412	0.273	0.180
1984	0.283	0.412	0.844	0.412	0.283
1985	0.180	0.273	0.412	0.857	0.436
1986	0.124	0.180	0.283	0.436	0.899
Truies					
1982	0.908	0.429	0.272	0.167	0.099
1983	0.429	0.866	0.405	0.262	0.167
1984	0.272	0.405	0.853	0.405	0.272
1985	0.167	0.262	0.405	0.866	0.429
1986	0.099	0.167	0.272	0.429	0.908
Bovins					
1982	0.914	0.438	0.264	0.135	0.061
1983	0.438	0.870	0.412	0.253	0.135
1984	0.264	0.412	0.856	0.412	0.264
1985	0.135	0.253	0.412	0.870	0.438
1986	0.061	0.135	0.264	0.438	0.914

La méthode des moindres carrés généralisés peut être appliquée à d'autres combinaisons d'estimateurs (selon le groupe de renouvellement et l'année) mais les résultats observés laissent croire qu'on y gagnerait peu.

5. CONCLUSIONS

L'estimateur composite qui est proposé dans cet article offre un moyen de combiner plusieurs estimateurs de cheptels. Cet estimateur utilise les valeurs des divers estimateurs à base aréolaire et à base multiple observées dans l'année pour laquelle on cherche à établir une estimation officielle et dans plusieurs années antérieures. La combinaison linéaire optimale des six estimateurs dans une année particulière a une variance qui est de 2 à 12% moins élevée que celle de l'estimateur à base multiple à segment pondéré. Si l'on fait intervenir les estimateurs des quatre autres années, on réduit de 1 ou de 2% de plus la variance de l'estimateur composite pour l'année courante. Les données requises pour calculer l'estimateur à base multiple à segment pondéré servent aussi à calculer les cinq autres estimateurs. L'étape la plus exigeante dans la construction de l'estimateur composite est d'estimer la matrice des covariances des estimateurs relatifs aux années pour lesquelles il existe des données d'échantillon. Comme les variances sont relativement fixes d'une année à l'autre, on peut calculer d'avance le vecteur de poids et l'appliquer aux estimations de l'année courante. Ensuite, calculer l'estimateur composite pour l'année d'estimation n'est qu'une formalité.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à un contrat de recherche (n° 53-319T-6-00073) passé avec le National Agricultural Statistics Service du Département de l'agriculture des E.-U. Les auteurs tiennent à remercier Ron Fecso et Vic Tolomeo pour l'aide qu'ils leur ont apportée. Ils remercient plus particulièrement Ron Fecso pour les commentaires qu'il a faits sur une version antérieure de l'article. Cette étude a été réalisée pendant que le premier auteur était en congé de formation à l'Université Iowa State.

Soit la matrice de corrélation pour les estimateurs de groupe de renouvellement, Z ,

$$W = \begin{pmatrix} W_0 & W_1 & W_2 & W_3 & W_4 \\ W_1 & W_0 & W_1 & W_2 & W_3 \\ W_2 & W_1 & W_0 & W_1 & W_2 \\ W_3 & W_2 & W_1 & W_0 & W_1 \\ W_4 & W_3 & W_2 & W_1 & W_0 \end{pmatrix}$$

où $W_0 = I_5$,

$$W_1 = \begin{pmatrix} 0 & \rho_1 & 0 & 0 & 0 \\ 0 & 0 & \rho_1 & 0 & 0 \\ 0 & 0 & 0 & \rho_1 & 0 \\ 0 & 0 & 0 & 0 & \rho_1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0 & 0 & \rho_2 & 0 & 0 \\ 0 & 0 & 0 & \rho_2 & 0 \\ 0 & 0 & 0 & 0 & \rho_2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$W_3 = \begin{pmatrix} 0 & 0 & 0 & \rho_3 & 0 \\ 0 & 0 & 0 & 0 & \rho_3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad W_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & \rho_4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Alors, l'estimateur à base aréolaire à segment pondéré par les moindres carrés généralisés est

$$\tilde{q} = (D'W^{-1}D)^{-1}D'W^{-1}Z,$$

où W est l'estimateur pour la matrice de corrélation W . La matrice des covariances de \tilde{q} est estimée par

$$Cov(\tilde{q}) \equiv (D' \tilde{\Sigma}^{-1} D)^{-1},$$

où $\tilde{\Sigma} \equiv 5W$ est la matrice des covariances de Z , dont les éléments sont tels que la variance estimée de l'estimateur à base aréolaire à segment pondéré est égale à un. Le tableau 6 donne les matrices des covariances estimées, $Cov(\tilde{q})$, pour les trois catégories de bétail. Nous constatons que pour 1986, les estimateurs calculés à l'aide d'estimations tirées des groupes de renouvellement surpassent d'environ 10% en efficacité les estimateurs à base aréolaire à segment pondéré. Les poids optimaux pour le vecteur des estimations de groupes de renouvellement sont

$$(D'W^{-1}D)^{-1}D'W^{-1}.$$

On peut obtenir les poids en question en s'adressant aux auteurs.

Tableau 5
Coefficients de corrélation estimés des estimateurs à base aréolaire à segment pondéré pour un même groupe de renouvellement

h	Porcs	Truies	Bovins
0	1.000	1.000	1.000
1	0.606	0.590	0.592
2	0.478	0.456	0.433
3	0.365	0.336	0.258
4	0.304	0.217	0.097

4.4 Estimation à l'aide des moyennes de groupes de renouvellement

En calculant les estimations de la Section 4.3, nous n'avons pas utilisé tous les renseignements dont nous disposons. Nous avons utilisé les estimateurs pour chaque année mais n'en avons pas extrait les éléments propres à chaque groupe de renouvellement. Dans cette section, nous allons construire un estimateur en nous servant des moyennes de groupes de renouvellement de l'estimateur à base aréolaire à segment pondéré. Nous posons l'hypothèse que la variance de l'estimateur est la même pour les cinq années. Suivant cette hypothèse, les coefficients de corrélation sont supposés dépendre uniquement du nombre d'années qui séparent les estimateurs en question. Soit ρ_h le coefficient de corrélation des estimateurs à base aréolaire à segment pondéré pour un même groupe de renouvellement observé à h années d'intervalle, $h = 0, 1, \dots, 4$. Les coefficients de corrélation estimés sont présentés dans le tableau 5 pour les trois cheptels. Ces coefficients de corrélation sont la moyenne des coefficients de corrélation estimés à partir des 5 - h groupes de renouvellement concernés. On compte en tout neuf groupes de renouvellement pour les cinq années.

Soit Z_{tj} l'estimateur à base aréolaire à segment pondéré pour le groupe de renouvellement j pour l'année t , où $j = 1, \dots, t + 1, \dots, t + 4$ et $t = 1, 2, \dots, 5$. Alors pour une année donnée t , nous supposons que Z_{tj} est un estimateur non biaisé du nombre total de têtes de bétail α_t . Nous savons qu'il peut exister un biais de renouvellement et que ce biais doit être estimé; toutefois nous n'en tiendrons pas compte ici. Le modèle est

$$Z_{tj} = \alpha_t + \epsilon_{tj}, \quad t = 1, 2, \dots, 5; \quad j = t, t + 1, \dots, t + 4, \tag{4.5}$$

où les erreurs, ϵ_{tj} , ont une moyenne nulle. En notation matricielle, le modèle (4.5) s'écrit

$$Z = D\alpha + \tilde{\epsilon},$$

où

$$Z' = (Z_{11}, Z_{12}, \dots, Z_{15}; Z_{22}, Z_{23}, \dots, Z_{26}; \dots; Z_{55}, Z_{56}, \dots, Z_{59}),$$
$$\tilde{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_5),$$

$D = I_5 \otimes 1_5$, I_5 étant la matrice unité d'ordre 5 et 1_5 le vecteur (5×1) dont tous les éléments égaient un.

4.3 Estimation du modèle

Etant donné la matrice des covariances estimée, V^* , nous estimons les paramètres du modèle (3.4) à l'aide de l'estimateur par les moindres carrés généralisés estimé (3.5). Les valeurs de l'estimateur composite, $\hat{\alpha}_T$, pour les cheptels de 1986 figurent à la première ligne du tableau 3. Celui-ci donne également les effets des six estimateurs, désignés par β_i dans le modèle (3.3). L'écart type estimé de l'estimateur composite est légèrement plus élevé que celui de l'estimateur à base multiple à segment pondéré. L'augmentation de variance vient de ce que la valeur de l'estimateur est établie à l'aide d'estimations d'échantillon et d'estimations officielles antérieures. Les sommes des carrés des résidus définies en (3.7) sont de 18.22, 15.38 et 24.59 pour les porcs, les truies et les bovins respectivement. Le nombre de degrés de liberté est de 20 puisque, pour chaque cheptel, il y a trente observations dans le V^* -vecteur et que dix paramètres sont estimés dans $\hat{\gamma}$. La somme des carrés des résidus ne dépasse en aucun cas 31.41, valeur qui correspond au 95-ième percentile pour la distribution du chi-carré à 20 degrés de liberté.

Les valeurs de l'estimateur composite qui figurent dans le tableau ci-dessus ont une erreur type comparable à celle des valeurs de l'estimateur à base multiple à segment pondéré, auquel correspond justement l'erreur type la plus faible (tableau 1). On s'attendrait donc que la combinaison linéaire optimale des six estimateurs pour une seule année attribue le poids le plus élevé à l'estimateur à base multiple à segment pondéré et c'est justement ce qu'on observe. Les poids à variance minimum pour les données d'une seule année sont calculés par la formule

$$(1'V^*_0 - 1'1) - 1'V^*_0 - 1,$$

où $1' = (1, 1, 1, 1, 1, 1, 1)$ et V^*_0 est la matrice des covariances des six estimateurs présentée dans le tableau 2 (voir les éléments diagonaux de (4.3)). Les poids optimaux et l'erreur type estimée de la combinaison optimale des six estimateurs figurent dans le tableau 4. Notons que la somme des poids fait un pour chaque cheptel. L'écart entre les erreurs types indiquées dans ce tableau et celles indiquées à la première ligne du tableau 3 est dû à l'estimation de niveau dans le calcul des estimations du tableau 3.

Tableau 4

Poids optimaux de six estimateurs pour une seule année.

Type d'inventaire					
Estimateurs			Porcs	Truies	Bovins
Estimateurs à base aérolaire					
Segment fermé					
			0.0541	-0.0152	0.0525
Segment ouvert					
			-0.0084	0.0152	0.0656
Segment pondéré					
			0.1463	0.1909	0.0909
Estimateurs à base multiple					
Segment fermé					
			0.1640	-0.0218	-0.0353
Segment ouvert					
			-0.0116	-0.0191	-0.0772
Segment pondéré					
			0.6556	0.8500	0.9035
Erreur type estimée de la combinaison optimale			0.94	0.95	0.99

Au lieu de considérer l'échantillon habituel, où vingt pour cent des segments sont renouvelés à chaque année, prenons un échantillon formé d'un ensemble de groupes de renouvellement qui ont été observés à chacune des cinq années. Pour cet échantillon, la matrice des covariances des six estimateurs pour les cinq années (exprimée en fonction des sous-matrices de (4.1)) s'écrit

$$(4.4) \quad \begin{pmatrix} V_{11} & \frac{4}{5}V_{12} & \frac{3}{5}V_{13} & \frac{2}{5}V_{14} & 5V_{15} \\ \frac{4}{5}V_{21} & V_{22} & \frac{4}{5}V_{23} & \frac{3}{5}V_{24} & \frac{2}{5}V_{25} \\ \frac{3}{5}V_{31} & \frac{4}{5}V_{32} & V_{33} & \frac{4}{5}V_{34} & \frac{3}{5}V_{35} \\ \frac{2}{5}V_{41} & \frac{3}{5}V_{42} & \frac{4}{5}V_{43} & V_{44} & \frac{4}{5}V_{45} \\ 5V_{51} & \frac{2}{5}V_{52} & \frac{3}{5}V_{53} & \frac{4}{5}V_{54} & V_{55} \end{pmatrix}$$

Les estimations directes des sous-matrices V_{ij} , qui sont calculées à partir de segments communs aux années i et j , donnent parfois une matrice des covariances (4.4) qui n'est pas définie positive. Cela peut arriver, par exemple, lorsque le groupe de renouvellement présente les cinq années comprend de très gros exploitants. Lorsqu'on intègre les hypothèses définies en (4.2) au processus d'estimation, les estimations de la matrice des covariances (4.4) sont définies positives pour les trois cheptels.

Tableau 3
Estimations composites pour les cheptels de 1986
et effets de divers estimateurs

	Porcs ¹	Truies ¹	Bovins ¹
Estimateur composite	18.84 (1.01)	18.06 (1.02)	16.43 (1.03)
Effets des estimateurs à base aréolaire			
Segment fermé	-1.13 (1.30)	-2.26 (1.36)	-0.21 (0.99)
Segment ouvert	0.26 (1.86)	-1.09 (1.78)	1.03 (1.45)
Segment pondéré	1.24 (1.14)	-0.94 (1.10)	-0.26 (0.80)
Effets des estimateurs à base multiple			
Segment fermé	-0.33 (0.66)	-1.86 (0.78)	0.04 (0.92)
Segment ouvert	-0.11 (0.75)	-1.82 (0.84)	1.40 (1.32)
Segment pondéré	0.19 (0.59)	-1.74 (0.59)	-0.31 (0.69)

¹ Les erreurs types figurent entre parenthèses.

Soit S^* la matrice diagonale 6×6 formée des racines carrées de la moyenne des variances estimées des six estimateurs pour les cinq années. Là encore, aux fins de la protection du secret statistique, on normalise les variances estimées de manière que la variance estimée de l'estimateur à base multiple à segment pondéré soit égale à 1.00. Alors, la matrice des covariances estimée pour les six estimateurs pour les cinq années est

$$V^* =$$

$$\begin{pmatrix} S^* & 0 & 0 & 0 & 0 & 0 \\ 0 & S^* & 0 & 0 & 0 & 0 \\ 0 & 0 & S^* & 0 & 0 & 0 \\ 0 & 0 & 0 & S^* & 0 & 0 \\ 0 & 0 & 0 & 0 & S^* & 0 \\ 0 & 0 & 0 & 0 & 0 & S^* \end{pmatrix} \begin{pmatrix} C_0^* & C_1^* & C_2^* & C_3^* & C_4^* & C_0^* \\ C_1^* & C_2^* & C_3^* & C_4^* & C_0^* & C_1^* \\ C_2^* & C_3^* & C_4^* & C_0^* & C_1^* & C_2^* \\ C_3^* & C_4^* & C_0^* & C_1^* & C_2^* & C_3^* \\ C_4^* & C_0^* & C_1^* & C_2^* & C_3^* & C_4^* \\ C_0^* & C_1^* & C_2^* & C_3^* & C_4^* & C_0^* \end{pmatrix} \begin{pmatrix} S^* & 0 & 0 & 0 & 0 & 0 \\ 0 & S^* & 0 & 0 & 0 & 0 \\ 0 & 0 & S^* & 0 & 0 & 0 \\ 0 & 0 & 0 & S^* & 0 & 0 \\ 0 & 0 & 0 & 0 & S^* & 0 \\ 0 & 0 & 0 & 0 & 0 & S^* \end{pmatrix}$$

(4.3)

Le tableau 2 donne les matrices des covariances estimées, $V_o^* = S^* C_o S^*$, pour les cheptels. On peut obtenir les valeurs estimées des quatre sous-matrices non diagonales uniques, $V_r^* \equiv S^* C_r S^*$, $r = 1, 2, 3, 4$, en s'adressant aux auteurs.

Tableau 2

Matrices des covariances estimées pour les six estimateurs de cheptels pour une année

Estimateurs à base aréolaire			Estimateurs à base multiple		
Segment fermé	Segment ouvert	Segment pondéré	Segment fermé	Segment ouvert	Segment pondéré

3.886	4.077	2.366	0.654	0.688	0.405
4.077	7.959	2.394	0.698	1.150	0.430
2.366	2.394	2.784	1.242	1.590	0.937
0.654	0.698	0.373	1.239	1.936	0.937
0.688	1.150	0.409	1.590	1.937	0.937
0.405	0.430	0.481	0.936	0.937	1.000

B. Truies

4.720	4.274	2.455	1.102	1.112	0.572
4.274	7.260	2.322	1.119	1.427	0.548
2.455	2.322	2.621	0.481	0.487	0.499
1.102	1.119	0.481	1.638	1.658	1.033
1.112	1.427	0.487	1.658	1.934	1.033
0.572	0.548	0.499	1.033	1.033	1.000

C. Bovins

2.355	1.951	1.141	1.853	1.655	0.907
1.951	5.527	1.014	1.652	4.418	0.912
1.141	1.014	1.321	0.913	0.891	0.925
1.853	1.652	0.913	1.910	1.756	0.992
1.655	4.418	0.891	1.756	4.310	1.017
0.907	0.912	0.925	0.992	1.017	1.000

où V_{ij} représente les sous-matrices de (4.1); $r = 0, 1, \dots, \max(5-t, 5-j)$ et $t, j = 1, 2, \dots, 5$.

Pour $t = j$ et $r = 0$, les hypothèses de (4.2) impliquent

$$V_{11} = V_{22} = V_{33} = V_{44} = V_{55} \equiv V_0.$$

Pour $t \neq j$, les hypothèses de (4.2) impliquent la relation suivante:

$$V_{12} = V_{23} = V_{34} = V_{45} \equiv V_1,$$

$$V_{13} = V_{24} = V_{35} \equiv V_2,$$

$$V_{14} = V_{25} \equiv V_3,$$

et

$$V_{15} \equiv V_4.$$

Ces hypothèses s'accordent assez bien avec les données. Une telle concordance était prévisible puisque la taille de l'échantillon varie peu au cours des cinq années et qu'il n'y a pas de variation notable des cheptels.

Nous estimons chacune des sous-matrices de (4.1) en faisant la moyenne des matrices des covariances estimées correspondantes établies à partir des segments communs. Pour cela, nous nous fondons sur les matrices de corrélation. Exprimons la matrice des covariances des totaux estimés définie en (4.1) par la formule

$$V = S C S,$$

où S est la matrice diagonale 30×30 des écarts types estimés des six estimateurs pour les cinq années et C est la matrice de corrélation 30×30 , partitionnée de la même manière que V définie en (4.1).

Nous construisons l'estimateur de la matrice de corrélation C en faisant la moyenne des sous-matrices estimées de C . À l'aide des segments communs à deux années, nous estimons la matrice des covariances des deux vecteurs de totaux estimés établis au moyen de ces segments en appliquant la formule habituelle pour l'échantillonnage en grappes stratifié. Nous transformons ensuite les matrices des covariances estimées en matrices de corrélation et les estimations ainsi obtenues sont appelées estimations directes. Soit

$$\begin{aligned} C_0 &= \left(\frac{5}{2}\right) \frac{1}{2} (C_{11} + C_{22} + C_{33} + C_{44} + C_{55}) \\ C_1 &= \left(\frac{5}{4}\right) \frac{1}{4} (C_{12} + C_{23} + C_{34} + C_{45}) \\ C_2 &= \left(\frac{5}{3}\right) \frac{1}{3} (C_{13} + C_{24} + C_{35}) \\ C_3 &= \left(\frac{5}{2}\right) \frac{1}{2} (C_{14} + C_{25}) \\ C_4 &= \left(\frac{5}{2}\right) C_{15}, \end{aligned}$$

où C_{ij} représente les matrices de corrélation estimées directement et fondées sur les segments communs. Les facteurs entre parenthèses représentent la portion des segments qui sont communs aux estimations. Cette fraction découle du plan d'échantillonnage avec renouvellement selon lequel vingt pour cent des segments de l'échantillon aréolaire sont renouvelés chaque année. En vertu de l'hypothèse d'indépendance, la corrélation entre les segments supprimés de l'échantillon et ceux qui y sont introduits est nulle.

Puisque les matrices de corrélation estimées C_{ij} ne sont pas symétriques lorsque $i \neq j$, nous appliquons l'hypothèse de la symétrie énoncée en (4.2) ($V'_{ij} = V_{ji}$) à la matrice des covarian-

ces estimée en posant

$$C'_r = \frac{1}{2} (C_r + C'_r), \quad r = 1, 2, 3, 4.$$

Le tableau 1 donne les estimations des cheptels pour 1986 ainsi que les écarts types estimés des estimateurs pour la même année. Chaque écart type indiqué est la racine carrée de la moyenne des variances estimées pour chacune des cinq années. Nous avons calculé les valeurs du tableau en fixant à 1.00 la valeur de l'écart type de l'estimateur à base multiple à segment pondéré pour tous les cheptels. Cette mesure a pour effet de faciliter la comparaison et respect du même coup les règles de protection du secret statistique.

Comme des études antérieures le laissaient prévoir (voir, par exemple, Nealon 1984), l'estimateur à base aréolaire à segment ouvert est le moins précis de tous en ce qui a trait aux cheptels. L'estimateur le plus précis est celui à base multiple, dans sa version à segment pondéré pour le domaine des non-répétoriés. Les estimateurs à base aréolaire à segment pondéré ont des coefficients de variation qui vont de 7 à 9% tandis que les estimateurs à base multiple à segment pondéré ont des coefficients de variation qui se situent entre 5,5 et 6,5%. Comme l'échantillon de listage est plus grand pour les éleveurs de porcs que pour les éleveurs de bovins, la précision des estimateurs à base multiple, relativement aux estimateurs à base aréolaire, est beaucoup plus grande pour le cheptel porcin que pour le cheptel bovin.

4.2 Estimation des matrices des covariances

L'estimation de la matrice des covariances pour les six estimateurs et les cinq années de données se fait en plusieurs étapes. La matrice des covariances pour le vecteur d'erreurs, *e*, dans l'équation (3.4) peut s'écrire de la façon suivante:

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} & V_{14} & V_{15} \\ V_{21} & V_{22} & V_{23} & V_{24} & V_{25} \\ V_{31} & V_{32} & V_{33} & V_{34} & V_{35} \\ V_{41} & V_{42} & V_{43} & V_{44} & V_{45} \\ V_{51} & V_{52} & V_{53} & V_{54} & V_{55} \end{pmatrix}$$

(4.1)

où, pour une catégorie particulière de cheptel, *V_{ij}* est la matrice 6 x 6 des covariances des six estimateurs pour l'année *i* et des six estimateurs pour l'année *j*. Compte tenu du plan de renouvellement utilisé, la covariance des estimateurs pour deux années est une fonction du nombre de groupes de renouvellement qui sont communs aux deux années. Soit *k* = |*i* - *j*| pour *k* = 0, 1, ..., 4. Alors, on peut estimer la matrice des covariances, *V_{ij}*, à partir des segments aréolaires des 5 - *k* groupes de renouvellement qui sont communs aux années *i* et *j*.

Nous estimons les éléments de la matrice des covariances (4.1) en posant quelques hypothèses additionnelles à propos de sa structure. Notre but premier est de comparer la précision des divers estimateurs et les hypothèses qui suivent ont justement pour effet de faciliter cette comparaison.

Nous allons supposer que les matrices des covariances pour les couples d'années à écarts égaux sont égales et symétriques. Autrement dit, nous supposons

$$V_{ij} = V_{i+j+r}$$

et

(4.2)

$$V_{ij} = V'_{ij}$$

Le répertoire des producteurs de porc dans l'Etat analysé est divisé en onze strates, qui sont définies en fonction du nombre total de porcs qu'élevaient les producteurs à un moment précis. Ainsi, il y a des strates pour les exploitations agricoles où il n'y a pas de bétail, où il y a du bétail à l'exclusion des porcs, où il y a de 1 à 99 porcs, de 100 à 199 porcs, . . . , au-delà de 6,000 porcs. En ce qui concerne les éleveurs de bovins, le répertoire à partir duquel est formé l'échantillon de l'enquête énumérative de juin contient les noms de très gros producteurs. On parle d'une liste de moins de 500 producteurs pour chacune des années étudiées. Le répertoire des éleveurs de bovins se divise en quatre strates. Trois d'entre elles sont définies en fonction du nombre total de bovins, c'est-à-dire entre 1,000 et 2,999 bovins, entre 3,000 et 9,999 et au-delà de 10,000. La quatrième strate regroupe les exploitants agricoles qui ont au moins 200 vaches laitières.

L'échantillon aréolaire de l'enquête énumérative de juin comptait en moyenne 2,350 exploitants agricoles (plus ou moins 60) pour chaque année étudiée. Quant à l'échantillon de listage, il comptait en moyenne 2,400 exploitants (plus ou moins 50) dans le cas des porcs et 70 exploitants (plus ou moins 35,5) dans le cas des bovins. Les données de ces échantillons ont permis d'estimer le nombre total de porcs, de truies et de bovins pour chacune des cinq années étudiées en utilisant les six estimateurs définis plus haut. Les estimations ont été calculées à l'aide de PC CARP, un programme d'ordinateur personnel servant à l'estimation de paramètres à partir d'un échantillon d'enquête (voir Fuller et coll. 1986 et Schnell et coll. 1988). Les estimateurs de la variance sont les estimateurs habituels de la variance pour un total estimé établi à l'aide d'un échantillon en grappes stratifié. Voir par exemple Cochran (1977).

Les données qui ont servi au calcul de la variance étaient considérées comme des données complètes même si dans quelques cas, il a fallu recourir à l'imputation pour pallier à la non-réponse. Comme les méthodes d'imputation ont recours largement aux données d'années antérieures dans le schéma de renouvellement, cette opération peut avoir pour effet de surestimer la corrélation d'une année à l'autre.

Tableau 1
Estimations de cheptels pour 1986

	Porcs	Truies	Bovins
Estimateurs à base aréolaire			
Segment fermé	18.42	15.78	15.27
Segment ouvert	(1.97)	(2.17)	(1.53)
Segment pondéré	21.11	18.24	18.74
	(2.82)	(2.69)	(2.35)
Estimateurs à base multiple			
Segment fermé	18.11	15.59	16.12
Segment ouvert	(1.11)	(1.28)	(1.38)
Segment pondéré	18.50	15.82	16.22
	(1.00)	(1.00)	(1.00)

d'une année à l'autre puisque les mêmes segments aréolaires servent aux enquêtes pendant plusieurs années suivant un plan d'échantillonnage avec renouvellement. Quant à l'échantillon de liste, on en forme un nouveau à chaque année. Les variances et les covariances des estimateurs pour n'importe quelle année donnée peuvent être estimées à l'aide des méthodes de sondage ordinaires. Comme on se sert du même estimateur (fondé sur un échantillon de liste) pour définir les trois estimateurs à base multiple pour une année donnée, la covariance de deux de ces trois estimateurs pour la même année aura une composante due à la variance de l'estimateur construit à l'aide de l'échantillon de liste. On peut estimer les covariances d'estimateurs pour des années différentes, $\text{Cov}(X_{it}, Y_{it})$, où $t \neq t'$, par des méthodes courantes en utilisant les segments d'échantillon communs aux deux années. Si l'on suppose que les variances et les covariances en N satisfont des fonctions particulières, on peut intégrer ces conditions dans la méthode d'estimation.

Etant donné un estimateur de la matrice des covariances, désigné par V^* , l'estimateur par les moindres carrés généralisés estimé du vecteur des paramètres $\tilde{\gamma}$ est

$$\tilde{\gamma} = (X'V^*V^*-1X)^{-1}(X'V^*-1Y^*) \tag{3.5}$$

La matrice des covariances de $\tilde{\gamma}$ est estimée par

$$\widehat{\text{Cov}}(\tilde{\gamma}) = (X'V^*V^*-1X)^{-1} \tag{3.6}$$

L'estimateur par les moindres carrés généralisés estimé, $\hat{\alpha}_T$, qui est le $(T-1)$ -ième élément de $\tilde{\gamma}$ peut être un estimateur composite du cheptel pour l'année T . On peut estimer la variance de cet estimateur par l'élément correspondant de la matrice des covariances estimée (3.6). De plus, les estimateurs par les moindres carrés généralisés estimés, $\hat{\alpha}_T + \beta_i$, $i = 1, 2, \dots, N$, sont des estimateurs à base multiple redressés pour l'année T qui reposent sur le modèle (3.4). On estime les variances de ces estimateurs en définissant les fonctions linéaires appropriées de la matrice des covariances estimée (3.6).

Si le modèle (3.4) est juste et que les erreurs aléatoires sont distribuées suivant une loi normale, alors la somme des carrés pondérée

$$\chi^2 = (Y^* - X\tilde{\gamma})'V^*-1(Y^* - X\tilde{\gamma}) \tag{3.7}$$

suit une distribution de chi carré avec comme paramètre $NT - K$. La somme des carrés des résidus pondérée obtenue à l'aide de la matrice des covariances estimée permet donc de juger de façon approximative de la validité du modèle (3.1).

4. RÉSULTATS EMPIRIQUES

4.1 Introduction

Dans les enquêtes énumératives de juin réalisées entre 1982 et 1986 par le Département de l'agriculture des E.-U., on a échantillonné en tout 298 segments aréolaires dans l'Etat analysé. Ces segments ont été prélevés suivant un plan d'échantillonnage avec renouvellement qui prévoyait un taux de substitution annuel d'environ vingt pour cent. Bien que le taux de substitution réel varie, nous construisons les estimateurs comme s'il s'agissait d'un taux de vingt pour cent exactement.

La base aréolaire pour l'Etat est constituée de onze strates: neuf strates de terrains agricoles, cultivés dans des proportions diverses, une strate de terrains semi-agricoles et une strate pour les terrains à vocation domiciliaire ou commerciale.

où α_t est le cheptel pour l'année t ;
 β_t est l'effet lié à l'estimateur i ; et
 e_{it} est une erreur aléatoire de moyenne nulle.

Les effets de l'estimateur, $\beta_1, \beta_2, \dots, \beta_N$, sont là pour indiquer que des estimateurs différents pourraient avoir des espérances mathématiques différentes à cause des erreurs non dues à l'échantillonnage. Le modèle (3.1) stipule que les effets de l'estimateur sont additifs et constants d'une année à l'autre. L'hypothèse des effets constants est une simple précision qui s'accorde avec les données.

Le modèle (3.1) est un modèle type d'analyse de variance à deux critères, dont on ne peut estimer les paramètres sans poser d'hypothèses additionnelles. Afin de définir les paramètres du modèle, nous posons la condition suivante: la moyenne des cheptels réels pour les $(T-1)$ premières années doit être égale à la moyenne des estimations officielles correspondantes produites par l'Agricultural Statistics Board. Cette condition s'écrit

$$(3.2) \qquad \sum_{t=1}^{T-1} \alpha_t = \sum_{t=1}^{T-1} a_t,$$

où a_t est l'estimation officielle pour l'année t . Ainsi, nous sommes sûrs que les estimations de cheptels seront du même ordre que les estimations officielles déjà produites. Nous jugeons cette condition raisonnable puisqu'on ne peut connaître les valeurs réelles de f ou de α_t et qu'on ne doit pas négliger le fait que les estimations sont sous forme de séries chronologiques. Etant donné la condition (3.2), nous pouvons exprimer le modèle linéaire (3.1) en fonction des paramètres, $\alpha_2, \alpha_3, \dots, \alpha_T$ et $\beta_1, \beta_2, \dots, \beta_N$, par la formule

$$Y_{it}^* = - \sum_{j=2}^{T-1} \alpha_j + \beta_i + e_{it}$$

$$(3.3) \qquad \text{ou } t = 2, 3, \dots, T \text{ et } Y_{it}^* \equiv Y_{it} - \sum_{j=1}^{T-1} a_j, \quad i = 1, 2, \dots, N.$$

En notation matricielle, le modèle s'écrit

$$(3.4) \qquad Y^* = X\tilde{\gamma} + e,$$

où $Y^* \equiv (Y_{11}^*, \dots, Y_{1N}^*, Y_{21}^*, \dots, Y_{2N}^*, \dots, Y_{T1}^*, \dots, Y_{TN}^*)'$;
 X est la matrice $(NT \times K)$ des variables auxiliaires rattachées au modèle (3.3), où $K = T - 1 + N$;
 $\tilde{\gamma} \equiv (\alpha_2, \alpha_3, \dots, \alpha_T, \beta_1, \beta_2, \dots, \beta_N)'$; et
 e est le vecteur à NT colonnes des erreurs aléatoires avec matrice des covariances V .

La matrice des covariances, V , est la matrice des covariances des erreurs d'échantillonnage e_{it} qui se rattachent aux diverses méthodes d'estimation. Les estimateurs Y_{it} , \dots , $t = 1, 2, \dots, T$; $i = 1, 2, \dots, N$, sont corrélés pour n importe quelle année donnée puisqu'ils reposent sur les mêmes segments aréolaires et le même échantillon de listage. Ils sont aussi corrélés

qui ont trait à toute l'activité agricole des exploitations dont des secteurs se trouvent dans le segment. On réexprime les données en fonction du secteur en multipliant les totaux calculés pour l'exploitation par la proportion de la superficie totale de l'exploitation qui est comprise dans le segment. La valeur d'une variable pour un segment pondéré est égale à la somme des valeurs calculées pour chaque secteur du segment. On détermine les estimateurs de totaux (à segment fermé, ouvert ou pondéré) en multipliant la valeur calculée pour un segment par le poids de segment correspondant (inverse de la probabilité de sélection du segment) et en faisant la somme des valeurs ainsi obtenues pour tous les segments d'échantillon et toutes les strates d'un Etat. Houseman (1975) et Nealon (1984) analysent ces trois estimateurs et en font la comparaison. Pour la plupart des variables dont on peut enregistrer facilement la valeur par secteur, l'estimateur à base areolaire à segment fermé est jugé plus efficace que l'estimateur à base areolaire à segment ouvert. Or, des variables comme les dépenses agricoles et le nombre d'animaux morts sont difficilement évaluable au niveau du secteur. L'estimateur à segment fermé est plutôt utilisé pour estimer les superficies cultivées au niveau national et sert aussi, avec d'autres estimateurs, à évaluer les cheptels de la plupart des Etats. Lorsqu'on peut associer facilement les valeurs des variables aux secteurs, on préfère en règle générale l'estimateur à segment fermé parce qu'on croit que l'exploitant agricole a moins de chances de se tromper lorsqu'il fournit des données sur les secteurs plutôt que sur l'exploitation en général.

L'estimateur à base areolaire à segment pondéré est généralement le plus efficace de tous. Il peut servir à estimer un total de population pour n'importe quelle variable agricole. Nealon (1984, p. 19) cite plusieurs études qui montrent que l'estimateur à segment pondéré est biaisé parce qu'on minimise souvent la taille des exploitations agricoles. Certaines étendues de terre boisée, de terre à pâturage et de terre en friche ainsi que certaines parties de la ferme seraient omises par l'exploitant. En conséquence, le rapport de la superficie du secteur à la superficie totale de l'exploitation sera trop élevé et l'estimateur à base areolaire à segment pondéré sera biaisé positivement.

Les estimateurs à base multiple utilisent des données d'échantillon qui proviennent d'au moins deux bases. En ce qui concerne l'estimation des cheptels, on dispose habituellement de deux bases: base areolaire et répertoire. Le répertoire est la liste des exploitants qui, à un moment donné, faisaient l'élevage des animaux en question. Ce répertoire est incomplet mais il renferme en général le nom d'un bon nombre des principaux exploitants. Pour estimer le cheptel porcin d'un Etat à l'aide d'un estimateur à base multiple, on additionne l'estimateur du total pour le répertoire, construit à l'aide de l'échantillon de listage, et l'estimateur du total pour le domaine des non-répertoriés (exploitants dont le nom ne figure pas dans le répertoire), construit à l'aide de l'échantillon areolaire. L'échantillon de listage et l'échantillon areolaire sont tenus pour indépendants l'un de l'autre. L'estimateur à base multiple sera différent selon que l'estimateur pour le domaine des non-répertoriés sera un estimateur à segment fermé, à segment ouvert ou à segment pondéré.

3. ESTIMATEUR COMPOSITE

Pour l'estimation des cheptels, nous proposons un estimateur composite construit suivant l'hypothèse que la relation entre les divers estimateurs qui le composent est définie par un modèle linéaire. Supposons que nous ayons N estimateurs d'un cheptel donné pour T années consécutives individuellement et qu'il existe des estimations officielles du Agricultural Statistics Board pour les $T-1$ premières années. Nous cherchons à construire un estimateur composite du cheptel à la T -ième année.

Supposons que Y_{it} représente l'estimateur i -ième pour l'année t , où $t = 1, 2, \dots, T$ et $i = 1, 2, \dots, N$. Nous supposons le modèle linéaire,

$$Y_{it} = \alpha_i + \beta_i + e_{it},$$

(3.1)

et les données pertinentes sont envoyées à l'Agribusiness Statistics Board (ASB) du NASS à Washington (D.C.). Lorsqu'il établit les estimations officielles, l'ASB considère les divers estimateurs, les recommandations du bureau de l'Etat, les données de l'industrie, de même que les états récapitulatifs et les bilans par région. De plus, l'ASB a recours à la construction de graphiques pour maintenir la continuité chronologique entre les sources de données. L'organisme doit faire en sorte que la somme des estimations officielles pour chaque Etat corresponde aux estimations officielles nationales.

Un inconvénient majeur de la méthode utilisée actuellement pour calculer l'estimation officielle est qu'il n'existe pas de mesure de précision pour cette estimation. En 1983, un groupe de planification à long terme du NASS a recommandé qu'on élabora une méthode objective pour créer un estimateur composite à partir des divers estimateurs fondés sur un échantillon probabiliste (voir Allen et coll. 1983). En 1984, on a recommandé qu'un estimateur composite soit soumis à l'attention de l'Agribusiness Statistics Board (voir Bynum et coll. 1985, p. 2). Le regroupement de données provenant d'échantillons distincts mais liés entre eux et la combinaison de plusieurs estimateurs font l'objet de recherches statistiques depuis de nombreuses années. Kuo (1986) fait état de quelques-unes de ces recherches. Par la même occasion, il considère un estimateur composite des cheptels fondé sur les données des enquêtes du Département de l'Agriculture des E.-U.

Dans cet article, nous analysons une méthode permettant de construire un estimateur composite pour le nombre de têtes de bétail. Pour construire un tel estimateur, nous servons des valeurs de plusieurs estimateurs de cheptels calculées pour un certain nombre d'années ainsi que des variances et des covariances de ces estimateurs pour ces années. En supposant qu'un modèle linéaire simple explique les rapports entre ces estimateurs, nous obtenons l'estimateur par les moindres carrés généralisés des cheptels pour la dernière année pour laquelle il existe des données d'échantillon. A cause de l'importance de la série chronologique des estimations, l'ensemble des estimateurs composites est assujéti à la condition suivante: la moyenne des estimations pour toutes les années précédant l'année courante doit être égale à la moyenne des estimations officielles correspondantes. Ainsi, on maintient une certaine correspondance entre la série chronologique et les estimations officielles antérieures. D'autres conditions du même genre peuvent être définies.

2. ESTIMATEURS À BASE AREOLAIRE ET À BASE MULTIPLE

Dans la partie de l'enquête énumérative de juin qui repose sur une base aréolaire, on forme un échantillon de segments circonscrits sur des cartes puis on relève le nom de tous les exploitants agricoles qui exercent une activité dans ces segments et on les interviewe. Les interviewers déterminent si l'exploitant qui exerce une activité agricole dans un segment demeure dans ce même segment. On désigne par le terme «secteur» un terrain (ou groupe de terrains) compris dans un segment d'échantillon qui correspond à un mode d'exploitation en particulier. Un secteur peut représenter une exploitation agricole complète ou une partie de celle-ci. L'interviewer recueille des données sur l'activité agricole pour chaque secteur d'un segment d'échantillon et s'informe notamment de la taille du secteur. Il recueille en outre des données sur toute l'activité agricole de chaque exploitant de l'échantillon. Ces données permettent de construire trois estimateurs de totaux que l'on appelle respectivement estimateurs à base aréolaire à segment fermé, à segment ouvert et à segment pondéré. Ces estimateurs se distinguent particulièrement l'un de l'autre par la façon dont les données agricoles sur lesquelles ils reposent se rattachent au segment.

L'estimateur à base aréolaire à segment fermé utilise des données concernant l'activité agricole dans chaque secteur d'un segment. L'estimateur à base aréolaire à segment ouvert utilise des données qui ont trait à toute l'activité agricole des exploitations dont le propriétaire habite dans le segment. Enfin, l'estimateur à base aréolaire à segment pondéré utilise des données

Estimation des cheptels à l'aide de plusieurs estimateurs à base de sondage aréolaire et à base de sondage multiple

GEORGE E. BATTESE, NANCY A. HASABELNABY et WAYNE A. FULLER¹

RÉSUMÉ

Les auteurs cherchent à estimer le cheptel porcin et le cheptel bovin d'un Etat à l'aide des données de l'enquête énumérative de juin, qui est réalisée par le National Agricultural Statistics Service du Département de l'agriculture des Etats-Unis. Six estimateurs peuvent être construits à l'aide de ces données. Trois d'entre eux reposent sur des données d'échantillons aréolaires et les trois autres réunissent des données tirées d'enquêtes avec échantillonnage sur liste et d'enquêtes à base aréolaire. Un plan d'échantillonnage avec renouvellement est utilisé pour la partie de l'enquête énumérative de juin qui repose sur une base aréolaire. À l'aide de données pour la période 1982-1986, les auteurs estiment les covariances des estimateurs pour diverses années. Ils proposent un estimateur composite pour établir le nombre de têtes de bétail. Ils déterminent cet estimateur en faisant une régression par moindres carrés généralisés du vecteur formé de divers estimateurs annuels par rapport à un ensemble approprié de variables auxiliaires. L'estimateur composite est censé produire des estimations qui sont du même ordre que les estimations officielles du Département de l'agriculture des E.-U.

MOTS CLÉS: Enquête énumérative de juin; échantillon avec renouvellement; estimateur composite; moindres carrés généralisés.

1. INTRODUCTION

Le National Agricultural Statistics Service (NASS) (anciennement le Statistical Reporting Service) du Département de l'agriculture des E.-U. réalise en juin de chaque année des enquêtes probabilistes (enquêtes énumératives de juin) qui visent à recueillir des données sur les activités des exploitations agricoles. Ces données sont indispensables pour établir les estimations officielles concernant le nombre de têtes de bétail, les superficies cultivées, les stocks de céréales, etc. pour chacun des Etats et le pays en général. Les unités d'échantillonnage des enquêtes agricoles proviennent de bases aréolaires et de répertoires.

Dans le cas d'un Etat, la base aréolaire est le territoire de cet Etat, stratifié suivant le mode d'exploitation du sol, c'est-à-dire selon le pourcentage de territoire consacré à la culture et selon qu'il s'agit d'une région principalement urbaine, d'une région boisée, d'une région comprenant surtout des lacs ou d'autres régions non agricoles. Les unités d'échantillonnage sont appelées en l'occurrence des «segments» et leur taille varie selon les Etats et les strates mais est environ un mille carré dans les régions rurales.

En ce qui a trait à l'estimation des cheptels, on prélève en plus des échantillons d'exploitants agricoles à partir de listes contenant les noms des exploitants qui élèvent la catégorie de bétail en question. Ces listes ou répertoires sont stratifiées selon la taille de l'exploitation. On combine les données des enquêtes à base aréolaire et des enquêtes avec échantillonnage sur liste pour obtenir des estimateurs à base multiple qui permettront d'estimer le cheptel d'un Etat. On peut construire divers estimateurs à partir de l'échantillon aréolaire et de l'échantillon de listage. Les statisticiens des bureaux du NASS situés dans les Etats calculent plusieurs estimations et proposent une estimation officielle pour le cheptel d'un Etat. Toutes les propositions

¹ George E. Battese, Department of Econometrics, University of New England, Armidale, N.S.W. 2351 Australie. Nancy A. Hasabelnaby et Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011, Etats-Unis.

- HAIJEK, J. (1971). Comment. Dans *Foundations of Statistical Inference*, Eds. V. P. Godambe et D. A. Sprott. Toronto: Holt, Rinehart, et Winston.
- HUMAN NUTRITION INFORMATION SERVICE (1985). *CSFII - Nationwide Food Consumption Survey Continuous Survey of Food Intake by Individuals: Women 19-50 Years and Their Children 1-5 Years, 1 Day*. NFCS, CSFII Report No. 85-1. Washington: United States Department of Agriculture.
- ISAKI, C. T., et FULLER, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTT, P. S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, 73, 485-491.
- KOTT, P. S. (1987). Estimating the conditional variance of a design consistent regression estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 486-491.
- PRASAD, N. G. N., et RAO, J. N. K. (1986). On the estimation of mean square error of small area predictions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 108-116.
- SÄRNDAAL, C. E. (1984). Design consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCOTT, A., et SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.
- SHAH, B. V. (1981). *SESUDAAAN: Standard Error Program for Computing of Standardized Rates from Sample Survey Data*. Research Triangle Park: Research Triangle Institute.
- STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. Dans *Small Area Statistics*, Eds. R. Platek, J. N. K. Rao, C. E. Särndal et M. P. Singh. New York: John Wiley and Sons.

Comme nous l'avons dit plus haut, une façon de choisir l'un ou l'autre des deux estimateurs $v(e_j)$ — $v(d_j)$ qui est indépendante du modèle d'enchaînement, est de dénombrer simplement les fois que δ_j et e_j est négatif. Cependant, l'estimateur $v(e_j)$ est instable et ne devrait pas servir en pratique à estimer l'erreur quadratique moyenne.

Tandis que les estimations de l'erreur quadratique moyenne de e_j ont instables, les $v(d_j)$ sont à peine meilleures. Au mieux, le nombre de «degrés de liberté» se rattachant à $v(d_j)$ est égal à la différence entre le nombre d'UPF et le nombre de strates dans J . En ce qui concerne l'échantillon de la CSFII, le nombre de degrés de liberté varie de 2 à 7.

Comme les statisticiens doivent de plus en plus indiquer les erreurs types estimées à côté des moyennes estimées qu'ils publient, il est urgent de trouver des estimateurs plus stables que $v(d_j)$ et $v(e_j)$. On pourrait, par exemple, ajuster $v(d_j)$ et $v(e_j)$, au moyen d'une fonction d'estimation de la variance, cet ajustement pouvant s'appliquer à l'un ou à l'autre estimateur ou aux deux à la fois. Toutefois, il s'agit là d'une méthode *ponctuelle* qui ne peut guère plus que donner des valeurs proches des valeurs estimées (entièrement dépendantes du modèle) des erreurs quadratiques moyennes de d_j et de e_j (voir Prasad et Rao 1986, pour une analyse probante de la question) en faisant la moyenne des effets du rejet du modèle.

Une solution intéressante serait de combiner les estimateurs (stables mais biaisés) de l'erreur quadratique moyenne fondés sur le modèle avec les estimateurs convergents selon le plan qui ont été définis dans cet article, un peu comme on le fait pour les moyennes avec e_j . Cependant, d'autres recherches s'imposent pour que nous sachions comment appliquer cette solution.

REMERCIEMENTS

L'auteur tient à remercier le Human Nutrition Information Service, qui a permis l'accès à sa base de données, ainsi que Joe Goldman, qui a aidé à constituer les séries de données qui ont servi à l'analyse empirique. Des remerciements sont également adressés à John Herbert et à deux arbitres anonymes pour leurs précieux commentaires sur des versions préliminaires de cet article.

BIBLIOGRAPHIE

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BREWER, K. R. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5, 93-105.

BREWER, K. R. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

FAY, R. E., et HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FULLER, W. A., et HARTER, R. M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*, Eds. R. Platek, J. N. K. Rao, C. E. Särndal et M. P. Singh. New York: John Wiley and Sons.

GHOSH, M., et MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.

GONZALEZ, M.E., et HORA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

de croire que les e_j sont effectivement de meilleures estimations que les d_j . De façon formelle, si nous considérons les 24 écarts d_j , comme indépendants d'un domaine à l'autre (ce qui n'est pas tout à fait le cas), l'hypothèse que les erreurs quadratiques moyennes réelles de modèle (ou de plan) de $v(e_j) - v(d_j)$, sont égales et que la variable aléatoire $v(e_j) - v(d_j)$ est aussi susceptible de prendre une valeur positive que négative ne tient plus du tout.

Si nous substituons e_j à d_j , l'erreur quadratique moyenne devrait diminuer d'environ 40,6 % (selon la formule $\sum \{v(e_j) - v(d_j)\} / \sum v(d_j)$). L'erreur-typique devrait pas conséquent diminuer de 22,9 %. Comme nous additionnons 24 variables aléatoires quasi indépendantes, nous avons beaucoup plus confiance dans cette estimation que dans n'importe quelle estimation particulière $v(e_j)$ (ou $v(d_j)$) dans les circonstances).

5. ANALYSE

Soit $n_j^* = 1 / \sum_{i=1}^{n_j} w_i^2$ la taille effective de l'échantillon dans le domaine j . Souignons que $n_j^* \leq n_j$, l'égalité étant vérifiée si et seulement si tous les poids d'échantillonnage dans j égaient $1/n_j$. Pour un rapport σ^2/δ^2 , connu, la seule chose qui distingue l'estimateur optimal défini ici $f_j(a^*, c^*)$, du meilleur estimateur linéaire sans biais de Scott et Smith (1969) est que $1/n_j^*$ remplace $1/n_j$ dans la formule qui sert à calculer α^* (équation (5)). Cette substitution a pour effet d'accroître w_{ji} lorsque les poids j ne sont pas tous égaux; en d'autres termes, cette substitution crée une dépendance plus forte par rapport aux données d'échantillon qui ne sont pas du domaine j . Cela est inévitable car lorsqu'on s'efforce d'obtenir un estimateur convergent selon le plan, on n'utilise pas l'échantillon du domaine j de la façon la plus efficace. Nous pourrions toujours pondérer cet échantillon d'une manière conforme à partir d'autres domaines en nous servant de poids d'échantillon pour calculer $\mu^*(L)$, mais cela ne ferait que réduire l'efficacité de modèle de l'estimateur sans améliorer aucune caractéristique axée sur le plan.

Par l'équation (7), nous sommes sûrs que L ne peut être inférieur à zéro. Cela implique que $\alpha^*(L)$ ne peut être plus grand que $\sum_{g=j} n_g^* / (\sum_{g=j} n_g^* + n_j^*)$. Si $\alpha^*(L)$ était égal à sa borne supérieure et que $n_j^* = n_j$, e_j se réduirait alors à la moyenne simple des y_{gi} pour tout l'échantillon. Cela a du sens car lorsque le modèle défini en (4) est exact et que $\sigma^2 = 0$, l'estimateur le plus efficace de $\mu + \gamma_j = \mu$ est la moyenne de l'échantillon global.

Par contre, si $n_j^* < n_j$ et $L = 0$, on calculera e_j en pondérant plus fortement les unités qui n'appartiennent pas au domaine j que celles qui y appartiennent, ce qui est peu rationnel. Une façon ponctuelle de contourner le problème serait de définir une borne supérieure de $1 - (n_j / \sum n_g)$ (ou moins) pour $\alpha^*(L)$. Une autre solution serait de renoncer à l'estimation pour un petit domaine lorsque la valeur de $\alpha^*(L)$ calculée selon la formule proposée plus haut, dépasse $1 - (n_j / \sum n_g)$. Pour que cela se réalise, il faudrait que L , la valeur estimée pour σ^2/δ^2 , soit très faible. Dans l'exemple empirique de la section précédente, la valeur de L se situait entre 0,03 et 0,06 et pourtant $\alpha^*(L)$ était encore bien en-deçà de $1 - (n_j / \sum n_g)$.

Le modèle complet défini par l'équation (4) peut s'avérer inexact de deux façons: ou bien le modèle des effets fixes dans chaque domaine (équation (1)) est erroné, ou bien le modèle d'enchaînement (équation (3)) l'est. Dans la réalité, les deux modèles sont susceptibles d'être erronés. Le modèle des effets fixes ne tient pas compte des effets de stratification ou de grappe ni des effets, aussi tenus soient-ils, de la présence de plusieurs femmes du même groupe d'âge dans un ménage. Dans l'un ou l'autre cas, ces effets ne devraient pas être notables. De plus, en introduisant des poids d'échantillonnage dans l'estimateur d_j et en faisant en sorte que les estimateurs de l'erreur quadratique moyenne soient convergents selon le plan, nous avons fait tout ce que nous pouvions pour parer au rejet du modèle des effets fixes.

Par ailleurs, le modèle d'enchaînement ne devrait pas nous inspirer une grande confiance. Ce modèle n'est guère plus qu'un artifice statistique qui, notamment, ne tient pas compte de la corrélation qui pourrait exister entre les quantités d'aliments consommés par des femmes qui vivent dans la même région mais dans des secteurs qui n'ont pas le même niveau d'urbanisation ou vice versa.

Tableau 1
Valeurs estimées pour les domaines, par groupe d'âge

Domaine	Taille de l'échantillon	Femmes 19-34 ans				
		d_j	e_j	$v(d_j)$	$v(e_j)$	$\alpha'(L)$
N - C	68	220.6	222.1	683.0	367.5	.233
N - B	95	195.7	203.1	568.8	367.8	.225
N - R	12	219.1	223.8	5266.7	-1349.5	.630
M - C	55	270.7	258.6	2021.5	1152.5	.251
M - B	107	277.2	267.8	625.8	509.6	.164
M - R	73	301.1	285.9	4027.1	2754.3	.187
S - C	66	212.4	215.7	3011.6	1700.1	.220
S - B	112	156.8	167.9	472.8	457.3	.146
S - R	81	117.0	139.3	592.0	868.9	.184
O - C	39	403.0	333.2	2064.2	5438.4	.364
O - B	74	205.0	209.6	1704.0	1018.3	.207
O - R	13	120.0	190.7	3533.5	3924.3	.652
Femmes 35-50 ans						

D'après Kott (1987),

$$v(d) = v^*(d) \text{ var}_e(d) / E_e[v^*(d)]$$

est à la fois un estimateur convergent selon le plan de l'erreur quadratique moyenne de d (suivant certaines conditions) et un estimateur non biaisé selon le modèle de la variance de modèle de d .

Le tableau 1 donne les valeurs de $n_j, d_j, \alpha'(L), e_j, v(d_j)$ et $v(e_j)$ qui ont été calculées pour les 12 domaines dans chacun des deux groupes d'âge (l'indice de domaine j figure de nouveau avec d_j et e_j). En utilisant l'équation (5), on obtient une valeur de L égale à 0.055 pour les femmes de 19 à 34 ans et à 0.037 pour les femmes de 35 à 50 ans. Ces résultats donnent à penser que les femmes d'un domaine ont peu de choses en commun si ce n'est qu'elles appartiennent au même groupe d'âge. Néanmoins $\alpha'(L)$ n'est supérieur à 0.5 que pour les cinq domaines (sur un total de 24) pour lesquels l'échantillon compte moins de 25 femmes.

La valeur estimée $v(e_j)$ est négative dans deux cas sur 24 et inférieure à $v(d_j)$ dans 18 cas sur 24, soit dans neuf cas pour chaque groupe d'âge. Ces derniers chiffres nous permettent

Avant de poursuivre, nous devons définir de nouvelles relations. Soit

$$x_{hk} = \sum_{i=1}^{n_{hk}} w_{hki},$$

$$z_{hk} = \sum_{i=1}^{n_{hk}} w_{hki}^2,$$

$$f_{hk} = \sum_{i=1}^{n_{hk}} w_{hki} (y_{hki} - d),$$

et

$$f_h = \sum_{k=1}^{K_h} f_{hk}/K_h.$$

En supposant que la taille de la population du domaine est suffisamment élevée pour ne pas être prise en considération (cette hypothèse permettant aussi d'affirmer avec une assez grande certitude qu'aucune personne n'a été échantillonnée deux fois), la variance de modèle de d est

$$\begin{aligned} \text{var}_e(d) &= \delta^2 \sum_h \sum_k z_{hk} \\ &= \delta^2 \sum_h \sum_k z_{hk}. \end{aligned}$$

L'estimateur SESUDAN (linéarisation) de l'erreur quadratique moyenne de plan de d est

$$v^*(d) = \sum_H (K_h/[K_h - 1]) \sum_{k=1}^{K_h} (f_{hk} - f_h)^2.$$

Après de nombreuses transformations, il est possible de montrer que l'espérance de modèle de cet estimateur est

$$E_e[v^*(d)] = \delta^2 \sum_h \sum_k z_{hk}$$

$$\begin{aligned} &- 2 \sum_h (K_h/[K_h - 1]) \left(\sum_{k=1}^{K_h} z_{hk} x_{hk} - \sum_{k=1}^{K_h} x_{hk}^2 \right) \\ &+ \left(\sum_h \sum_k z_{hk} \right) \left(\sum_{k=1}^{K_h} x_{hk}^2 - \sum_{k=1}^{K_h} x_{hk}/K_h \right). \end{aligned}$$

Les statisticiens ont souvent une confiance beaucoup plus grande dans le modèle de base (équation 1) que dans le modèle d'enchaînement (équation 3), surtout lorsque ce dernier s'accompagne de l'hypothèse que les variances (δ_g) sont les mêmes d'un domaine à l'autre. Il est donc rassurant de savoir que l'on peut estimer la précision de e_j sans recourir à l'équation (3) ou exiger que les δ_g soient égales.

Malheureusement, $v(e_j)$ est instable et peut même être négatif lorsque $\alpha'(L)$ est supérieur à 0.5. Néanmoins, une simple comparaison des valeurs relatives de $v(d_j')$ et de $v(e_j')$ pour les m domaines ($j = 1, \dots, m$) représente une méthode robuste pour choisir l'un ou l'autre des deux estimateurs, d_j et e_j .

4. EXEMPLE EMPIRIQUE

Dans le cadre de la Continuing Survey of Food Intakes by Individuals (CSFII), le Human Nutrition Information Service (HNIS) a recueilli des données sur la quantité de nourriture consommée quotidiennement par des femmes de 19 à 50 ans en 1985; pour cela, l'organisme a procédé à un échantillonnage à plusieurs degrés stratifié. Les données recueillies visaient 60 groupes d'aliments et 27 éléments nutritifs. Voir Human Nutrition Information Service (1985) pour plus de détails sur l'enquête et le plan de sondage.

Nous allons nous borner ici à estimer la quantité moyenne de lait et de produits laitiers (1 des 60 groupes d'aliments) consommés par des femmes âgées de 19 à 34 ans et de 35 à 50 ans dans douze domaines s'excluant mutuellement. Ces domaines sont définis en fonction de deux critères de classification: la région (Nord-Est, Middle-West, Sud et Ouest) et le niveau d'urbanisation (centre-ville, banlieue, secteur non métropolitain). En ce qui concerne la consommation moyenne par groupe d'aliments, HNIS a publié des données distinctes pour les deux groupes d'âge au niveau national seulement. Pour ce qui est de la consommation moyenne d'éléments nutritifs, des données ont été publiées pour chaque groupe d'âge selon la région et le niveau d'urbanisation mais non en fonction des deux critères à la fois.

Le plan de sondage de la CSFII prévoyait un échantillon à plusieurs degrés stratifié indépendant pour chacun des domaines. Tout d'abord, les unités primaires d'échantillonnage (villes ou municipalités) ont été choisies au moyen d'un échantillonnage avec remise avec probabilité proportionnelle à la taille; ensuite, nous avons prélevé un sous-échantillon aléatoire de segments aréolaires duquel a été tiré un sous-échantillon aléatoire de ménages, plus petit. Nous avons ensuite procédé à un autre sous-échantillonnage. En effet, lorsqu'un ménage de l'échantillon de la CSFII comptait plusieurs femmes du même groupe d'âge, nous en choisissons une aléatoirement.

Pour chaque groupe, d_j défini en (2) représente l'estimateur classique (fondé sur un plan) de la moyenne d'un domaine. Le programme SESUDAAN (Shah 1980) permet de calculer des estimateurs convergents selon le plan pour tous les d_j et leurs erreurs moyennes de plan ($\sqrt{\text{EQM}(d_j)}$). Cependant, lorsque ces estimateurs sont élevés au carré, ils ne sont pas nécessairement des estimateurs non biaisés de la variance de d_j selon le modèle défini en (1). Afin de vérifier cela, nous allons nous en tenir non seulement à un seul groupe d'âge mais aussi à un seul domaine et nous allons supprimer l'indice inférieur j . Désignons les strates par $h = 1, \dots, H$ les unités primaires d'échantillonnage (U.P.E.) dans, $k = 1, \dots, K_h$ et les femmes choisies dans hk par $i = 1, \dots, n_{hk}$. L'estimateur pour la consommation moyenne est

$$d = \sum_{h=1}^H \sum_{k=1}^{K_h} \sum_{i=1}^{n_{hk}} w_{hki} y_{hki}.$$

Si le modèle (4) est juste et que tous les $\delta_j^2 = \delta^2 > 0$, alors L doit être positif pour une valeur de m suffisamment grande. Même si le modèle n'est pas valide, pour autant que L a une borne inférieure positive $|\mu^*(L)|$ est bornée et $n_j \sum_{i=1}^{n_j} w_{ji}^2$ est bornée lorsque n_j (mais non m) prend une valeur arbitrairement élevée, alors e_j est convergent selon le plan lorsque d_j l'est car

$$\lim_{n_j \rightarrow \infty} \mu^*(L) = 0,$$

de sorte que e_j converge vers d_j qui est convergent selon le plan.

3. ERREUR QUADRATIQUE MOYENNE DE MODÈLE ET DE PLAN

Selon certains plans de sondage, l'estimateur de la variance de plan de d_j est aussi un estimateur non biaisé selon le modèle de d_j pris comme estimateur de y_{jp} selon le modèle de base (pour simplifier l'exposé, nous omettrons désormais les mots «pris comme estimateur de y_{jp} »). Cependant, on doit souvent prendre en compte un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j (à supposer, comme nous le ferons, qu'il existe une telle erreur). Cela est particulièrement vrai lorsque $\sum_{k=1}^{K-1} d_j^{k-1} \neq N_j$. Kott (1987) montre comment (lorsque cela s'impose) on peut redresser cet estimateur de manière à en faire simultanément un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j et un estimateur non biaisé de la variance de d_j selon le modèle de base. Désignons cet «estimateur de variance» redressé par $v(d_j)$.

Nous pouvons maintenant parler des erreurs quadratiques moyennes de modèle et de plan de l'estimateur des effets aléatoires, e_j . Bien qu'il faille supposer que les δ_j^2 étaient tous égaux pour déterminer e_j , il n'est pas nécessaire de faire de même lorsqu'on évalue la précision de e_j . De fait, il n'est même pas nécessaire de supposer que le modèle d'enchaînement défini par l'équation (3) est valide! Il suffit de supposer que m est assez grand pour que L puisse être considéré comme (virtuellement) indépendant des unités du domaine de j . Par ailleurs, on peut redéfinir L en retirant les unités du domaine j des sommes qui figurent dans le membre de droite de l'équation (7).

$$D'une façon ou d'une autre, $E_e[(d_j - y_{jp})(y_{jp} - \mu^*(L))] = 0$. En conséquence,
$$E_e[\{d_j - \mu^*(L)\}^2] = \text{var}_e(d_j - y_{jp}) + E_e[\{y_{jp} - \mu^*(L)\}^2].$$$$

Il est maintenant facile de montrer que selon le modèle de base défini en (1),

$$v(e_j) = [1 - 2\alpha^*(L)] v(d_j) + [\alpha^*(L)]^2 [d_j - \mu^*(L)]^2$$

est un estimateur non biaisé de l'erreur quadratique moyenne de modèle de e_j étant donné L et $\mu^*(L)$. Puisque $\alpha^*(L)$ est asymptotiquement nul lorsque n_j tend vers l'infini, $v(e_j)$ est aussi un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de e_j lorsque $v(d_j)$ est un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j .

Il n'est pas nécessaire que L converge vers σ^2/δ^2 ou que $\mu^*(L)$ converge vers μ pour que $v(e_j)$ possède les propriétés décrites ci-dessus. De fait, il n'est pas nécessaire de donner quel-que interprétation que ce soit aux limites de L et $\mu^*(L)$ puisque les propriétés en question ont été définies indépendamment du modèle exprimé par l'équation (3).

Si nous supposons que tous les δ^2_{ε} sont égaux à δ^2 , il est facile de montrer, par la méthode des multiplicateurs de Lagrange, que les valeurs de α^2 , et de c^g qui minimisent la variance de modèle de $f_j(\alpha, c) - y_j^p$ sont

$$\alpha^* = \frac{\sum_{j=1}^I w_j^2 f_j^i + \sum_{g=1}^G c_{*2}^g / n_g + (1 + \sum_{g=1}^G c_{*2}^g) (\sigma^2 / \delta^2)}{\sum_{j=1}^I w_j^2 f_j^i - 1 / N_j} \tag{5}$$

et

$$c_{*}^g = \frac{\sum_{h=1}^h [(\sigma^2 / \delta^2) + n_{-1}^h] - 1}{[(\sigma^2 / \delta^2) + n_{-1}^g] - 1}, \text{ pour } g \neq j. \tag{6}$$

Dans la pratique, σ^2 et δ^2 sont rarement connues. Ghosh et Meeden (1986) ont proposé d'estimer le rapport σ^2 / δ^2 à partir de l'échantillon et d'une manière qui soit conforme au modèle (lorsque $m \rightarrow \infty$) par la formule

$$L = \max \left\{ 0, \left[\frac{\sum_{g=1}^G n_g (y_{gS}^2 - y_S^2) / (m - 1)}{\sum_{g=1}^G \sum_{i=1}^I (y_{gi}^2 - y_{gS}^2) / (n - m)} - 1 \right] \right\} \tag{7}$$

où

$$y_S^2 = \sum n_g y_{gS}^2 / n$$

et

$$n = \sum n_g.$$

Posons $\alpha'(L)$ et $c'(L)$ comme les membres de droite des équations (5) et (6) respectivement, où L remplace σ^2 / δ^2 . Ensuite, désignons

$$e_j = f_j[a'(L), c'(L)]$$

comme l'estimateur des effets aléatoires, où μ dans $e_j = f_j(\cdot, \cdot)$ équivaut à $\mu'(L) = \sum c_g'(L) y_{gS}$. Lorsque m augmente, e_j se confond de plus en plus avec $f_j(a^*, c^*)$.

Isaki et Fuller (1982) exposent les conditions suffisantes pour que d_j soit convergent selon le plan et il l'est effectivement selon la plupart des plans de sondage d'usage courant, à l'exception notamment de l'échantillonnage systématique effectué à partir d'une liste pré-établie (voir Kott 1986). Au lieu de la convergence selon le plan, on parle souvent, pour un estimateur, de la propriété d'être *asymptotiquement non biaisé selon le plan* (Brewer 1979). L'estimateur d_j est toujours asymptotiquement non biaisé selon le plan. L'inconvénient de cet estimateur d_j est qu'il peut ne pas être très efficace lorsque n_j est faible. Pour éliminer cet inconvénient, on peut notamment «tirer avantage» des autres domaines en considérant le paramètre fixe θ_j comme la réalisation d'une variable aléatoire, qui satisfait le modèle d'enchaînement:

$$\theta_j = \mu + \tau_j, \tag{3}$$

où $E(\tau_j) = 0$, et $E(\tau_j \tau_g) = \sigma^2$ lorsque $j = g$ et 0 dans le cas contraire. C'est ce qu'on appelle parfois le «modèle des effets aléatoires» car l'effet j , θ_j , qui était jusque-là fixe, est maintenant considéré comme une variable aléatoire. En combinant les équations (1) et (3), on obtient le modèle des composantes de variance en forme abrégée:

$$y_{ji} = \mu + \tau_j + \epsilon_{ji}, \tag{4}$$

De nombreux analystes utilisent dès le départ l'équation (4). Nous avons fait la distinction entre le modèle de base et le modèle d'enchaînement pour mieux illustrer le fait que les analystes prêtent souvent plus de crédibilité au modèle de base (surtout lorsqu'on suppose dans le modèle d'enchaînement que tous les $\delta^2_g = \delta^2_2$, comme ce sera le cas bientôt). N'importe quel estimateur de la forme:

$$f_j(\alpha, c) = (1 - a)d_j + \alpha \hat{\mu},$$

où

$$c = (c_1, \dots, c_{j-1}, 0, c_{j+1}, \dots, c_m),$$

$$\hat{\mu} = \sum_m c_g \bar{y}_{gs},$$

$$\bar{y}_{gs} = \sum_{u=1}^I y_{gu} / n_g,$$

et

$$\sum_m c_g = 1$$

est non biaisé selon le modèle (4). (Nota: bien que les variables c et $\hat{\mu}$ dépendent du domaine j , on a supprimé les indices correspondants pour plus de simplicité.)

Dans la section 2, nous définissons un estimateur d'effets aléatoires convergent selon le plan pour la moyenne de population d'un petit domaine. Dans la section suivante, nous présentons un estimateur robuste (mais instable) pour les erreurs quadratiques moyennes de modèle et de plan de l'estimateur pour un petit domaine. L'estimateur est robuste en ce sens qu'il ne dépend pas du modèle fragile, mais nécessaire, qui relie les petits domaines entre eux. Enfin, nous présentons un exemple empirique dans la section 4 et une analyse dans la section 5.

2. L'ESTIMATEUR

Définissons tout d'abord le modèle *de base* (ou modèle des effets fixes):

$$(1) \quad y_{gi} = \theta_g + \epsilon_{gi},$$

où les ϵ_{gi} sont des variables aléatoires non corrélées de moyenne nulle et de variance $\text{var}(\epsilon_{gi}) = \delta_g^2$. L'indice inférieur désigne une unité appartenant au domaine g . N_g unités de la population se trouvent dans le domaine g et il y a m domaines. Prenons un domaine particulier j . Le problème consiste à estimer la moyenne de domaine:

$$\bar{y}_{jp} = \sum_{N_j}^{i=1} y_{ji}/N_j.$$

Soit p_{ji} la probabilité d'échantillonnage de l'unité ji et n_j le nombre d'unités prélevées dans le domaine j . Nous savons tous que si nous utilisons une méthode d'estimation linéaire non biaisée selon le plan et efficace selon le modèle pour \bar{y}_{jp} , p_{ji} sera égale à n_j/N_j et l'estimateur sera égal à $\sum_{i=1}^{n_j} y_{ji}/n_j$, où les unités sont identifiées à nouveau de manière que $j1, \dots, jn_j$ se trouvent dans l'échantillon.

Malheureusement, il arrive souvent en pratique que l'on doive estimer une moyenne de domaine à l'aide d'un échantillon qui n'a pas été nécessairement constitué à cette fin. Par conséquent, il se peut que les probabilités d'échantillonnage dans le domaine j ne soient pas toutes égales à n_j/N_j . Un estimateur très utilisé dans les circonstances est

$$(2) \quad d_j = \sum_{n_j}^{i=1} w_{ji} y_{ji},$$

où

$$w_{ji} = p_{ji}^{-1} / \sum_{n_j}^{k=1} p_{jk}^{-1},$$

désigne le poids d'échantillonnage de l'unité ji . Cet estimateur a été proposé par Brewer (1963) et Hajek (1971).

L'estimateur d_j est de toute évidence non biaisé selon le modèle (1), en ce sens que $E_\epsilon(d_j - \bar{y}_{jp}) = 0$. Il est aussi *convergent* selon de nombreux plans de sondage, c'est-à-dire

$$\begin{aligned} \text{plim}_\pi(d_j - \bar{y}_{jp}) &= 0, \\ n_j &\rightarrow \infty \end{aligned}$$

où π désigne l'espace-probabilité issu du processus d'échantillonnage aléatoire plutôt que du modèle (1).

Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires

PHILLIP S. KOTTI¹

RÉSUMÉ

Dans cet article, l'auteur utilise un modèle des effets aléatoires pour construire un estimateur convergent selon le plan pour un petit domaine. Il évalue ensuite l'erreur quadratique moyenne de cet estimateur sans supposer que la composante d'effet aléatoire du modèle est juste. À l'aide des données d'une enquête par sondage complexe, l'auteur montre comment cette méthode d'estimation de l'erreur quadratique moyenne, bien que probablement trop incertaine pour être appliquée directement, peut servir à déterminer si l'estimateur pour petits domaines proposé ici est supérieur à l'estimateur classique fondé sur un plan.

MOTS CLÉS : Population finie; modèle; erreur quadratique moyenne; convergent selon le plan; randomisation.

1. INTRODUCTION

Supposons que nous avons un échantillon probabiliste de valeurs unitaires et que l'on nous demande d'estimer la moyenne d'un petit domaine inclus dans la population visée par l'échantillon. Scott et Smith (1969) ont défini à cette fin un estimateur bayésien et ont montré qu'il était possible de construire cet estimateur essentiellement à l'aide de deux critères : l'absence de biais et la variance minimum. Nous utiliserons ici cette approche, qui est appelée parfois modèle des effets aléatoires ou modèle des composantes de variance.

La plupart des auteurs qui se sont penchés sur l'estimation pour petit domaine, outre Scott et Smith, (notamment, Fay et Herriot 1979; Battese et Fuller 1971; Ghosh et Meeden 1986; Prasad et Rao 1986; Fuller et Harter 1987; et Stroud 1987) supposent que le plan de sondage est non informatif et donc « neutre ». On pose la même hypothèse pour les estimateurs synthétiques de moyennes de petits domaines, dont il n'est aucunement question dans cet article (pour des exemples de ces estimateurs, voir Gonzalez et Hora, 1978).

Il faut supposer qu'un plan de sondage non informatif dissimule probablement l'apport le plus important de la randomisation à l'inférence statistique. Comme la plupart des modèles statistiques utilisés en inférence pour population finie sont soit erronés ou (au mieux) incomplets, il est souhaitable qu'une méthode d'estimation ait la propriété suivante: si l'échantillon est suffisamment grand, l'estimateur devrait approcher presque à coup sûr la valeur du paramètre qu'il vise à estimer, quel que soit le « véritable » modèle. Ce souhait trouve sa pleine expression dans le critère de convergence selon le plan, défini par Isaki et Fuller (1982).

La convergence selon le plan est une propriété asymptotique. On doit donc souvent définir par hypothèse un ou plusieurs modèles lorsque vient le temps de choisir parmi diverses méthodes d'estimation convergente selon le plan. Cela est particulièrement vrai dans le cas de l'estimation pour un petit domaine, où l'échantillon peut être particulièrement petit et le plan de sondage peut échapper à tout contrôle. Néanmoins, le fait de concentrer son attention sur des estimateurs convergents selon le plan offre une garantie minimale mais réelle en ce qui a trait à la validité du modèle. C'est pourquoi Särndal (1984) s'est surtout intéressé aux estimateurs convergents selon le plan pour petit domaine et c'est ce que nous proposons de faire dans cet article.

¹ Phillip S. Kotti, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, D.C., 20250, E.-U.

Souvent les estimations publiées qui ont trait à des enquêtes périodiques ne concernent que le dernier échantillon et n'exploitent donc pas de corrélations avec des estimations ayant trait à des périodes antérieures. D'autre part, il n'est pas rare que les économistes et d'autres spécialistes des sciences sociales ne tiennent aucun compte de l'erreur d'échantillonnage quand ils utilisent ces estimations dans leurs modèles de séries chronologiques. Binder et Dick montrent comment il peut être tenu compte de l'erreur d'échantillonnage dans ces modèles. Pour les lecteurs pour qui ce domaine est nouveau, les auteurs passent brièvement en revue les travaux effectués sur la question et donnent d'abondantes indications bibliographiques.

Battese, Hasabelnaby et Fuller étudient une méthode de construction d'un estimateur composite du stock de bétail. Ils utilisent un modèle linéaire pour intégrer sur plusieurs années six types d'estimateurs de l'Enquête énumérative de juin du Département de l'agriculture des États-Unis. Les résultats empiriques montrent qu'il y a une amélioration de la variance avec la combinaison linéaire optimale des six estimateurs pour une année donnée et que cette variance s'améliore encore si les estimateurs des autres années entrent en ligne de compte.

Berhel examine la répartition optimale dans les enquêtes à objectifs multiples. Il montre la sensibilité de la répartition optimale aux changements dans les contraintes de variance. L'auteur obtient des résultats qui peuvent être utilisés pour déterminer s'il est possible de réduire sensiblement les coûts d'une enquête en permettant une faible augmentation de certaines variances. Il présente également un algorithme d'itération conçu pour résoudre le problème de l'optimisation. Bruning et Hu font une comparaison de l'enquête-mémoire et de l'enquête-journal. Ils commencent par passer en revue les études où sont comparées les deux méthodes. Le corps de l'article traite d'une expérience qui a été faite pour définir quelle relation il y a entre certains facteurs démographiques et les méthodes de collecte. Les conclusions des auteurs confirment les résultats d'études antérieures mais soulèvent également la très réelle possibilité de problèmes de mesures associées à l'enquête-mémoire.

Lerneshow et Stroh examinent l'assurance de la qualité par échantillonnage comme moyen de réduire la taille de l'échantillon qui est nécessaire pour déterminer si l'état de santé d'une population répond à certaines normes. L'exemple choisi par les auteurs est celui de la couverture vaccinale des enfants dans les pays en développement. La méthode d'échantillonnage consiste à utiliser un échantillon initial pour tester l'hypothèse d'un taux de vaccination acceptable par strate. Les strates pour lesquelles le résultat n'est pas suffisamment concluant sont soumises à un nouvel échantillonnage.

Le rédacteur en chef

Dans ce numéro

Ce numéro de **Techniques d'enquête** contient une section spéciale sur l'utilisation statistique des données administratives. Les cinq articles formant cette section traitent de sujets divers, des questions touchant l'élaboration de politiques jusqu'au traitement des données.

Avec l'utilisation croissante des dossiers administratifs, il y a de plus en plus d'organismes statistiques qui font appel aux méthodes probabilistes d'appariement ou de couplage des registres. La plupart des applications utilisent la méthode décrite par Fellegi et Sunter (1969). Winkler examine l'importance d'une hypothèse d'indépendance qui est habituellement employée dans les applications du modèle Fellegi-Sunter parce qu'elle permet de beaucoup simplifier les calculs. Étudiant un problème d'appariement de listes d'entreprises, l'auteur cherche à déterminer quels changements peuvent être faits quand l'hypothèse d'indépendance n'est pas valide. L'article de Redfern traite d'un aspect de l'utilisation statistique des données administratives qui a beaucoup d'importance pour les organismes statistiques: l'utilisation des dossiers administratifs comme source de données de recensement traditionnel par questionnaire pour puiser ses données dans les dossiers administratifs. Dans trois autres pays d'Europe, certaines données jusqu'à obtenues au moyen d'un questionnaire viennent désormais directement de sources administratives. L'auteur étudie en détail la situation au Royaume-Uni. Il conclut que la résistance du public à ce qui pourrait être une violation de la vie privée, de même que l'idéologie politique et le manque de ressources, y sont des obstacles à l'intégration de renseignements administratifs de sources diverses dans un registre central de la population. L'auteur admet que les considérations politiques seront toujours le facteur primordial dans toute discussion relative à un registre de la population, mais il n'en estime pas moins que les statisticiens ont le devoir de faire connaître leur opinion.

Jonas et Hanczaryk observent que le rôle des données administratives au U.S. Bureau of the Census est devenu plus important au fil des ans. On a reconnu avant les recensements économiques de 1987 la nécessité d'un système global de gestion de la qualité qui permette de résoudre les problèmes du traitement de très grandes quantités de données. Le système mis au point prévoit l'utilisation extensive de micro-ordinateurs en vue de réduire les coûts.

Moore et Marquis décrivent une application des données administratives à l'évaluation d'estimations d'enquêtes. À l'aide de méthodes de couplage des enregistrements, des données de l'Enquête sur le revenu et la participation aux programmes effectuée par le U.S. Bureau of the Census ont été appariées aux dossiers administratifs pour cinq programmes fédéraux et quatre programmes d'états. L'analyse de cet ensemble de données ne fait que commencer. L'objet de cette analyse est de quantifier les effets des erreurs de mesure et d'utiliser les valeurs ainsi obtenues pour élaborer de meilleurs plans de sondage.

Statistique Canada procède actuellement à une restructuration de son programme d'enquêtes économiques. Un des éléments clés de cette restructuration est le remaniement du registre central d'entités économiques, qui servira de base de sondage pour les enquêtes économiques. L'article de Clark et Lussier expose les concepts et les méthodes qui serviront à l'établissement et à la mise à jour des profils d'entités économiques; il décrit également le rôle des données administratives dans cette tâche. Après une étude de simulation, l'article explore certaines questions ayant trait à l'établissement des profils.

Dans le premier article de ce numéro, Kott met au point un estimateur pour petits domaines qui répond au critère de convergence selon le plan proposé par Isaki et Fuller (1982). Il évalue l'erreur quadratique moyenne de cet estimateur. Au moyen d'un exemple empirique, Kott montre que l'estimateur de l'erreur quadratique moyenne peut être utilisé pour déterminer le choix entre l'estimateur proposé pour petits domaines et l'estimateur classique fondé sur le plan.

TABLE DES MATIÈRES

1	Dans ce numéro	P.S. KOTT
3	Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires ..	G.E. BATTESE, N.A. HASABELNABY et W.A. FULLER
15	Estimation des cheptels à l'aide de plusieurs estimateurs à base de sondage aréolaire et à base de sondage multiple	D.A. BINDER et J.P. DICK
31	Enquêtes répétées - Modélisation et estimation	J. BETHEL
49	Répartition de l'échantillon dans les enquêtes à plusieurs variables	E.R. BRUNING et M.Y. HU
61	Rôle des facteurs démographiques dans l'analyse de la précision de la déclaration des dépenses de consommation dans l'enquête-mémoire et dans l'enquête-journal ..	S. LEMESHOW et G. STROH JR.
73	Assurance de la qualité par échantillonnage pour l'évaluation des paramètres de santé dans les pays en voie de développement	Section Spéciale - Les utilisations statistiques des données administratives
		P. REDFERN
85	L'expérience européenne relative à l'utilisation des données administratives pour recenser la population: questions d'ordre politique	W.E. WINKLER
	Méthodes permettant de tenir compte de l'absence d'indépendance dans une application du modèle d'appariement des enregistrements de Fellegi-Sunter	J.R. JONAS et P.S. HANCZARYK
123	Contrôle automatisé de la qualité des données provenant de dossiers administratifs	J.C. MOORE et K.H. MARQUIS
133	Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes	C. CLARK et R. LUSSIER
151	Utilisation de données administratives pour l'établissement des profils initiaux et ultérieures des entités économiques	

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *U. of Western Ontario*

L. Biggert, *Université de Florence*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of*

Management and Budget

D. Holt, *University of Southampton*

K.M. Wolter, *A.C. Nielsen, U.S.A.*

V. Tremblay, *Statplus, Montréal*

F.T. Schuren, *U.S. Internal Revenue Service*

C.E. Särndal, *Université de Montréal*

I. Sande, *Bell Communications Research, U.S.A.*

D.B. Rubin, *Harvard University*

J.N.K. Rao, *Carleton University*

W.M. Podeschl, *Statistique Canada*

M.N. Murthy, *Applied Statistics Centre, India*

G. Kalton, *University of Michigan*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, 4^e étage, Édifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 30,00\$ par année au Canada, et de 35,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6. Un prix réduit, soit 16,00\$ (E.-U.) (20,00\$ Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

JUIN 1989

Publication autorisée par le ministre de
l'Expansion industrielle régionale

©Ministre des Approvisionnements
et Services Canada 1989

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable d'une
autorisation écrite du Groupe des programmes et produits
d'édition, agent intermédiaire aux permis, administration
des droits d'auteur de la Couronne, Centre d'édition
du gouvernement du Canada, Ottawa, Canada KIA 0S9.

Septembre 1989

Prix: Canada, \$30.00 par année
Autres pays, \$35.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 15, n° 1

ISSN 0714-0045

Ottawa

TECHNIQUES D'Échantillonnage

UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 15, NUMÉRO 1
JUN 1989

Canada



12
- 001



Statistics
Canada

Statistique
Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 15, NUMBER 2
DECEMBER 1989

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1989

Published under the authority of
the Minister of Regional Industrial Expansion

©Minister of Supply
and Services Canada 1990

Extracts from this publication may be reproduced for individual use without permission provided the source is fully acknowledged. However, reproduction of this publication in whole or in part for purposes of resale or redistribution requires written permission from the Programs and Publishing Products Group, Acting Permissions Officer, Crown Copyright Administration, Canadian Government Publishing Centre, Ottawa, Canada K1A 0S9

March 1990

Price: Canada, \$30.00 a year
Other Countries, \$35.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 15, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, <i>Statistics Canada</i>	M.N. Murthy, <i>Applied Statistics Centre, India</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	W.M. Podehl, <i>Statistics Canada</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
W.A. Fuller, <i>Iowa State University</i>	C.E. Särndal, <i>University of Montreal</i>
J.F. Gentleman, <i>Statistics Canada</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
D. Holt, <i>University of Southampton</i>	V. Tremblay, <i>Statplus, Montreal</i>
G. Kalton, <i>University of Michigan</i>	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

J. Gambino, J.-L. Tambay and A. Thériège, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$30.00 per year in Canada, \$35.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US \$16.00 (\$20.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 15, Number 2, December 1989

CONTENTS

In This Issue	159
Special Section – Data Analysis	
J.N.K. RAO, S. KUMAR and G. ROBERTS	
Analysis of Sample Survey Data Involving Categorical Response Variables: Methods and Software	161
Comments: R.E. FAY	180
C.J. SKINNER	181
E.A. MOLINA	183
Response: Authors	185
D.R. THOMAS	
Simultaneous Confidence Intervals for Proportions Under Cluster Sampling	187
J.G. MOREL	
Logistic Regression Under Complex Survey Designs	203
<hr/>	
L.A. FRANKLIN	
Randomized Response Sampling from Dichotomous Populations with Continuous Randomization	225
B. MacGIBBON and T.J. TOMBERLIN	
Small Area Estimates of Proportions Via Empirical Bayes Techniques	237
A. SUNTER	
Updating Size Measures in a PPSWOR Design	253
R.B.P. VERMA and R. RABY	
The Use of Administrative Records for Estimating Population in Canada	261
D.A. SWANSON	
Confidence Intervals for Postcensal Population Estimates: A Case Study for Local Areas	271
Acknowledgements	281

In This Issue

The risks involved in using standard statistical methods for the analysis of data from surveys with complex designs are becoming well-known. The special topic section in this issue contains three papers which provide guidance for the analysis of categorical data from such surveys. Tim Holt's efforts were instrumental in putting this section together.

The paper by Rao, Kumar and Roberts, which is the first discussion paper published in Survey Methodology, reviews developments in the analysis of cross-classified categorical data, extends them, and applies them to data from two large, complex surveys. The authors also briefly discuss computational issues. Comments by Fay, Skinner and Molina and a reply by Rao, *et al.* follow the paper.

Thomas describes a Monte Carlo study used to investigate several methods of obtaining simultaneous confidence intervals for proportions under a two-stage clustered design. He shows that some methods behave poorly, with actual coverage rates quite different from the nominal ones. Thomas concludes with guidelines on the choice of methods to use in practice.

The final paper in the section on data analysis for complex surveys, by Morel, deals with logistic regression. Using the results of a Monte Carlo study, he shows that for small samples, a modified Taylorization method for estimating a covariance matrix results in smaller biases than the usual delta method.

The bibliography by Nathan on randomized response which appeared in the previous issue of Survey Methodology attests to the large amount of research which has been devoted to the subject. In this issue, Franklin develops another approach to the randomized response model for sampling from dichotomous populations. The model is general in that it permits the use of randomization from a continuous distribution and multiple trials per respondent. Special attention is given to the case of randomization using the normal distribution function.

MacGibbon and Tomberlin examine the problem of small area estimation with complex survey designs. Their empirical Bayes estimator is a compromise between the highly variable but unbiased classical estimator and the more stable but potentially highly biased synthetic estimator.

A method of updating a PPSWOR sample which attempts to retain the same sample of primary sampling units is presented by Sunter. The method differs from earlier ones proposed by Kish and Scott (1971) and Fellegi (1963) in that it is valid for any sample size and does not require enumeration of all possible samples. The method is of particular importance for multistage survey samples which must be updated, but for which the cost of introducing new PSUs may be high.

Revenue Canada tax files and Family Allowance files are used in Canada to provide population estimates for provinces in non-census years. Verma and Raby examine the consistency of the estimates derived from these two sources. A comparison with the 1986 Census counts is also made.

Swanson presents a method of obtaining confidence intervals for post-censal population estimates. He shows that a Wilcoxon test can be used to determine if a change in model, due to post-censal structural changes, is required. Using empirical data, Swanson shows that ignoring such a change leads to confidence intervals whose coverage is lower than expected.

Analysis of Sample Survey Data Involving Categorical Response Variables: Methods and Software

J.N.K. RAO, S. KUMAR, and G. ROBERTS¹

ABSTRACT

During the past 10 years or so, rapid progress has been made in the development of statistical methods of analysing survey data that take account of the complexity of survey design. This progress has been particularly evident in the analysis of cross-classified count data. Developments in this area have included weighted least squares estimation of generalized linear models and associated Wald tests of goodness of fit and subhypotheses, corrections to standard chi-squared or likelihood ratio tests under loglinear models or logistic regression models involving a binary response variable, and jackknifed chisquared tests. This paper illustrates the use of various extensions of these methods on data from complex surveys. The method of Scott, Rao and Thomas (1989) for weighted regression involving singular covariance matrices is applied to data from the Canada Health Survey (1978-79). Methods for logistic regression models are extended to Box-Cox models involving power transformations of cell odds ratios, and their use is illustrated on data from the Canadian Labour Force Survey. Methods for testing equality of parameters in two logistic regression models, corresponding to two time points, are applied to data from the Canadian Labour Force Survey. Finally, a general class of polytomous response models is studied, and corrected chi-squared tests are applied to data from the Canada Health Survey (1978-79). Software to implement these methods using the SAS facilities on a main frame computer is briefly described.

KEY WORDS: Corrections to chi-squared tests; Logistic regression; Power transformations; Wald tests; Weighted least squares.

1. INTRODUCTION

Standard statistical methods, based on the assumption of independent identically distributed observations, are being used extensively by researchers in the social and health sciences, and in other subject matter areas. These methods have also been implemented in standard statistical packages, including SPSSX, BMDP, SAS and GLIM. In practice, however, much data are obtained from complex sample surveys involving clustering and stratification, so that the application of standard methods to these data without some adjustment for survey design can lead to erroneous inferences. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the complexity of the sample design is ignored in the analysis of data. Moreover, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, e.g., residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods, and emphasized the need for new methods that take proper account of the complexity of survey design. During the past 10 years or so, rapid progress has been made in the development of such methods, particularly for analysing cross-classified count data. This paper will focus on the analysis of

¹ J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario; S. Kumar and G. Roberts, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario.

count data, but it should be noted that important results on other types of analyses have also been obtained: Regression analysis (Fuller 1975; Nathan and Holt 1980; Pfefferman and Nathan 1981; Scott and Holt 1982), principal component analysis (Skinner, Holmes and Smith 1986), factor analysis (Fuller 1986), logistic regression involving continuous covariates (Binder 1983).

Rao and Scott (1984) have made a systematic study of the impact of survey design on standard Pearson chi-squared or likelihood ratio tests for multiway tables of counts, under hierarchical log-linear models. They have also obtained simple first order corrections to standard tests which can be computed from published tables that include "design effects" for cell estimates and marginal totals, thus facilitating secondary analyses from published reports (see also Gross 1984; Bedrick 1983; Rao and Scott 1987). These first order corrections take account of the design in the sense that the actual type I error rates of tests based on the corrected statistics are closer to nominal levels, compared to the standard tests which could have greatly inflated type I error rates. More accurate second order corrections, based on the Satterthwaite approximation to a weighted sum of independent χ^2 variables, were also developed by Rao and Scott (1984), but these tests require the knowledge of a full estimated covariance matrix of cell estimates. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975), and the jackknifed chi-squared tests (Fay 1985), all requiring either the full estimated covariance matrix or access to cluster-level data. Fay (1985) and Thomas and Rao (1987) have shown that the Wald statistic, although asymptotically correct, can become highly unstable as the number of cells in the multiway table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level. On the other hand, Fay's jackknife tests and the Rao-Scott corrections have performed well under quite general conditions. In some cases, the instability in the Wald statistic may be remedied by collapsing the table according to eigenvectors associated with the nonnegligible eigenvalues of the estimated covariance matrix adjusted for singularities caused by linear constraints on the probabilities, as proposed by Singh (1985); see also Singh and Kumar (1986).

Roberts, Rao and Kumar (1987) assumed a logistic regression model for the cell (domain) proportions associated with a binary response variable, and obtained first order corrections to standard chi-squared and likelihood ratio tests of goodness-of-fit and nested hypotheses. Simple upper bounds to first order corrections, depending only on the design effects of cell response proportions, were also obtained to facilitate secondary analyses from published tables. Scott (1986) proposed an alternative method which uses standard tests on transformed data derived from the original data and the cell design effects. Roberts, Rao and Kumar (1987) also provided second order corrections to standard tests, but these require access to a full estimated covariance matrix of cell response proportions. Diagnostics for detecting outliers and influential points were developed as well, again taking the survey design into account.

The primary purpose of this paper is to present various extensions of the previous methods and illustrate their use on data from large-scale surveys, including the Canada Health Survey (1978-1979) and the Canadian Labour Force Survey. It is assumed, throughout the paper, that the user has access to a full estimated covariance matrix of cell estimates. In Section 2, weighted least squares (WLS) estimators of the parameters of generalized linear models having singular covariance matrices, caused by linear constraints on the probabilities (or proportions), are presented. Associated Wald tests of goodness-of-fit and of subhypotheses are also provided. A smoothed version of the WLS estimators, and associated Wald tests of subhypotheses are given as well. These methods should be used only when the number of cells in a table is small and/or the number of sample clusters in the survey design is relatively large.

The methods for logistic regression models are extended, in Section 3, to Box-Cox models involving power transformations of cell odds ratios. These models, which include the logistic regression model as a special case, could provide significantly better fits than the logistic regression models, as demonstrated by Guerrero and Johnson (1982) in the context of binomial proportions.

Methods for testing equality of parameters in two logit models, corresponding to two different time periods, are given in Section 4. If the hypothesis of equality is accepted, one could obtain "smoothed" estimates of cell proportions for the current period that are more efficient than the corresponding smoothed estimates based only on the current period data.

Section 5 gives an extension of the type of results obtained for logistic regression models to a general class of polytomous response models. The special case of McCullagh's (1980) ordered response model is studied in detail.

Finally, an account of the software for implementing the above methodology is given in Section 6.

2. WEIGHTED LEAST SQUARES ESTIMATORS AND WALD TESTS

The approach of Koch, Freeman and Freeman (1975) is designed to estimate the parameters of generalized linear models of the form $g^*(p) = X^*\beta^*$, using a sample estimate, \hat{p} , of the population cell probabilities denoted by a T -vector p , and a consistent estimate of $\text{cov}(\hat{p}) = V_p$ (say). In this method, the asymptotic covariance matrix of the u -vector $g^*(p)$ is assumed to be nonsingular ($u < T$); however, many models, including the traditional loglinear model, are of the form $g(p) = X\beta$, where $g(p)$ is a T -vector with a singular asymptotic covariance matrix, and X is a $T \times r$ full rank matrix of known constants. It is possible to reduce the latter models to the nonsingular form $g^*(p) = X^*\beta^*$, as done by Grizzle and Williams (1972) for the loglinear model, but Scott, Rao and Thomas (1989) have developed the following unified approach for singular models, by appealing to the optimal theory for linear models having singular covariance matrices.

The cell probabilities p and \hat{p} are subject to linear constraints of the form $K'p = \pi$ and $K'\hat{p} = \pi$, where K is a $T \times L$ full rank matrix of known constants and π is an L -vector of known constants π_i ($L < T$). As a result, the covariance matrix of \hat{p} will be singular. For example, in the case of stratified sampling with complex sample designs within strata, we can write $K = I_L \otimes 1_m$, $\pi_i = n_i/n$ ($i = 1, \dots, L$) and $p = (p_{11} \dots p_{1m}; \dots; p_{L1} \dots p_{Lm})'$ with $p_{ij} = (n_i/n)\tilde{p}_{ij}$, where \tilde{p}_{ij} is the j -th category probability within the i -th stratum ($\sum_j \tilde{p}_{ij} = 1$; $i = 1, \dots, L$; $j = 1, \dots, m$), n_i is the sample size from the i -th stratum, $\sum n_i = n$, 1_m is a m -vector of 1's, I_L is the identity matrix of order L and \otimes denotes the Kronecker product.

Assume that $X\beta$ can be written as $X_0\beta_0 + X_1\beta_1$, where X_0 is a $T \times L$ matrix such that $K'H^{-1}X_0$ is nonsingular and where $H = (\partial g/\partial p)'$ is the $T \times T$ matrix of partial derivatives of $g(p)$. In particular, X_0 can be taken as K if the constraint matrix K is included in X , as frequently assumed. Since restrictions on p imply constraints on the parameters β , β_0 can be determined exactly from the constraints, for a given β_1 .

Weighted least squares estimators

The model may be written as

$$\hat{g} = g(\hat{p}) = X\beta + \delta \tag{2.1}$$

where δ is the error vector with $P \lim \delta = 0$, and \hat{g} has a singular asymptotic covariance matrix $V_g = H V_p H'$ which is consistently estimated as $\hat{V}_g = \hat{H} \hat{V}_p \hat{H}'$, assuming that \hat{V}_p is a consistent estimator of V_p . Here $\hat{H} = H(\hat{p})$. Scott, Rao and Thomas (1989) derived an asymptotically best linear unbiased estimator (ABLUE) of β_1 as

$$\hat{\beta}_1 = (\tilde{X}_1' \hat{M} \tilde{X}_1)^{-1} \tilde{X}_1' \hat{M} \hat{g}, \quad (2.2)$$

where

$$\hat{M} = (\hat{V}_g + X_0 X_0')^{-1} \quad (2.3)$$

is a nonsingular generalized inverse of \hat{V}_g , and

$$\tilde{X}_1 = [I - X_0 X_0' \hat{M}] X_1. \quad (2.4)$$

A consistent estimator of the asymptotic covariance matrix of $\hat{\beta}_1$ is given by

$$\text{est cov}(\hat{\beta}_1) = (\tilde{X}_1' \hat{M} \tilde{X}_1)^{-1}. \quad (2.5)$$

Wald tests

Letting $\hat{\beta} = (X' \hat{M} X)^{-1} X' \hat{M} \hat{g} = (\hat{\beta}_0', \hat{\beta}_1')'$, a Wald test of goodness of fit of the model (2.1) is given by

$$W = (\hat{g} - X \hat{\beta})' \hat{M} (\hat{g} - X \hat{\beta}) \quad (2.6)$$

which is distributed asymptotically as a χ^2 variable with $T - r$ degrees of freedom (d.f.). The model is considered tenable at the α -level if $W > \chi_{T-r}^2(\alpha)$, the upper α -point of χ^2 with $T - r$ d.f..

Given the model (2.1), tests of linear hypotheses on the model parameters β_1 can also be obtained. A Wald test of the linear hypothesis $C_1 \beta_1 = c_1$ is given by

$$W_1 = (C_1 \hat{\beta}_1 - c_1)' [C_1 \text{est cov}(\hat{\beta}_1) C_1']^{-1} (C_1 \hat{\beta}_1 - c_1) \quad (2.7)$$

which is distributed asymptotically as a χ^2 variable with h d.f., where C_1 is a $h \times (r - L)$ full rank matrix of known constants ($h < r - L$), and c_1 is a h -vector of known constants. The hypothesis is rejected at the α -level if $W_1 > \chi_h^2(\alpha)$, the upper α -point of χ^2 with h d.f. Note that β_0 should not be included in the linear hypothesis since it is fixed by the design constraints $K'p = K'g^{-1}(X\beta) = \pi$.

Smoothed version of ABLUE and associated Wald tests

We can also obtain a smoothed version of ABLUE of β_1 , say β_1^* , using iteration, as follows:

$$\check{\beta}_{t+1} = \check{\beta}_t + (X' M_t X)^{-1} X' M_t H_t (\hat{p} - p_t), \quad t = 0, 1, 2, \dots \quad (2.8)$$

with starting values $M_0 = \hat{M}$, $\check{\beta}_0 = (X' \hat{M} X)^{-1} X' \hat{M} \hat{g} = \hat{\beta}$, $H_0 = H(\hat{\beta})$ and $p_0 = p(\hat{\beta})$. Further, $M_t = (\hat{V}_{g_t} + X_0 X_0')^{-1}$ with $\hat{V}_{g_t} = H_t \hat{V}_p H_t'$, $H_t = H(\check{\beta}_t)$ and $p_t = p(\check{\beta}_t)$, $t \geq 1$. At convergence, we get $\beta^* = (\beta_0^*, \beta_1^{*'})'$ as the solution of the following equations:

$$X' M(\beta) H(\beta) (\hat{p} - p(\beta)) = 0. \quad (2.9)$$

Equations (2.9) reduce to quasilielihood equations (McCullagh 1983) when V_p is proportional to $V(p)$, a known function of p . Here, the dependence on β is made explicit by writing $p = p(\beta)$, $H = H(\beta)$ and $M = V_g + X_0X_0' = M(\beta)$. The smoothed estimate β^* also satisfies the constraints $K'p = K'g^{-1}(X\beta) = \pi$, unlike $\hat{\beta}$. The asymptotic covariance matrices of β_1^* and $\hat{\beta}_1$ are identical, but β_1^* might perform better in small samples.

Given the model (2.1), an alternate Wald test of the hypothesis $C_1\beta_1 = c_1$ is given by

$$W_1^* = (C_1\beta_1^* - c_1)' [C_1 \text{ est cov} (\beta_1^*) C_1']^{-1} (C_1\beta_1^* - c_1) \tag{2.10}$$

which is distributed asymptotically as a χ^2 with h d.f., where

$$\text{est cov} (\beta_1^*) = (X_1^{*'} M^* X_1^*)^{-1}, \tag{2.11}$$

and $X_1^* = [I - X_0X_0'M^*]X_1$, $M^* = (V_g^* + X_0X_0')^{-1}$ with $V_g^* = H^*\hat{V}_pH^{*'}$ and $H^* = H(\beta^*)$.

Example

The previous results were applied to a two-way table from the Canada Health Survey (1978-79). This survey was designed to provide reliable information on the health of Canadians. The information collected was made up of an interview component for the whole sample and a physical measures component for a subsample. A complex multistage design involving stratification and clustering was employed, and the estimates of cell totals or proportions were subjected to post-stratification on age-sex, to improve their efficiency. The reader is referred to Hidiroglou and Rao (1987) for a description of the survey and the procedures used for estimating cell counts, proportions, and their estimated variances and covariances. For the physical measures component, a collapsed stratum technique for variance estimation was employed since a single primary sampling unit was selected in some of the strata.

Table 1 gives the estimated proportions, \hat{p}_{ij} , derived from the physical measures component in a cross-classification of fitness level (recommended = 1, minimal acceptable = 2, below acceptable or screened out = 3) and type of cigarette smoker (regular = 1, occasional = 2, never = 3). The estimated covariance matrix of the \hat{p}_{ij} , \hat{V}_p , can be obtained from the authors.

Since both the variables in Table 1 are ordinal, we considered the following loglinear model with linear \times linear interaction:

$$\log p_{ij} = \tilde{u} + u_{1(i)} + u_{2(j)} + \gamma(v_i - \bar{v})(w_j - \bar{w}), \quad i = 1,2,3 \quad j = 1,2,3 \tag{2.12}$$

Table 1
Estimated Cell Proportions in a 3 \times 3 Table (Canada Level):
Type of Cigarette Smoker \times Fitness Level (Sample Size $n = 2505$)
Ages 15-64

Type of cigarette smokers	Fitness Level		
	1	2	3
1	0.22005	0.14951	0.16998
2	0.02301	0.00962	0.01146
3	0.20329	0.09933	0.11374

subject to side constraints $\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0$, where v_i and w_j are known scores with means \bar{v} and \bar{w} respectively. For simplicity, equidistant scores were taken: $u_i = 1,2,3$; $v_j = 1,2,3$. The model (2.12) is of the form $g(p) = X_0\beta_0 + X_1\beta_1$ with $g_{ij}(p) = \log p_{ij}$, $X_0 = K = 1_9$, a 9×1 vector of 1's, $\beta_0 = \bar{u}$, $\beta_1 = (u_{1(1)}, u_{1(2)}, u_{2(1)}, u_{2(2)}, \gamma)'$, and

$$X_1' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Noting that $\hat{H} = \text{diag}(\hat{p}_{ij}^{-1}, i = 1,2,3; j = 1,2,3)$, the Wald test of goodness-of-fit of the model (2.12) can be computed from (2.6), using the proportions \hat{p}_{ij} in Table 1 and the estimated covariance matrix, \hat{V}_p . We obtain

$$W = 3.59$$

which is not significant at the 5% level compared to $\chi^2_{T-r}(0.05) = \chi^2_3(0.05) = 7.81$ (note that $T = 9, r = 6$). The Wald statistic W is likely to be stable in this example since the number of cells $T (= 9)$ is small relative to the number of sample clusters $(= 50)$.

We can also conduct a test of independence, *i.e.* $\gamma = 0$, given the model (2.12), using W_1 given by (2.7) or W_1^* , based on the smoothed estimates β_1^* , given by (2.10). Noting that $C_1 = (0, \dots, 0, 1)$, $c_1 = 0$, we obtain

$$W_1 = 8.23, \quad W_1^* = 8.75,$$

both larger than $\chi^2_1(0.01) = 6.63$, the upper 1% point of χ^2 with 1 d.f. The nested hypothesis of independence is therefore not tenable.

Accepting the model (2.12), we obtain the following values of weighted least squares estimates, $\hat{\beta}_1$, and smoothed estimates, β^* :

$$\begin{aligned} \hat{\beta}_1 &= (0.912, -1.550, 0.339, -0.255, -0.086)' \\ \beta_0^* &= -2.665, \quad \beta_1^* = (0.917, -1.568, 0.344, -0.262, 0.087)'. \end{aligned}$$

The estimate β^* can also be used to produce smoothed estimates of the p_{ij} , $p_{ij}^* = p_{ij}(\beta^*)$, which satisfy the constraint $\sum \sum p_{ij}(\beta^*) = 1$.

3. BOX-COX TRANSFORMATION MODELS

Logistic regression models are extensively used for the analysis of variation in the estimated proportions associated with a binary response variable. Suppose that the population of interest is partitioned into I cells according to the levels of one or more factors. Let P_i be the population response proportion in the i -th cell. Then a logistic regression model for the proportions $P_i = F_i(\beta)$ is given by

$$\log\{F_i/(1 - F_i)\} = x_i'\beta, \quad i = 1, \dots, I, \tag{3.1}$$

where $x_i = (x_{i1}, \dots, x_{is})'$ is an s -vector of known constants derived from the factor levels with $x_{i1} = 1$, and β is an s -vector of unknown parameters.

Guerrero and Johnson (1982) extended the applicability of logistic regression models by introducing an additional parameter, λ , through a Box-Cox power transformation of the odds ratios $F_i/(1 - F_i)$. Their model is given by

$$v_i(\lambda) = \{F_i/(1 - F_i)\}^{(\lambda)} = x_i' \beta, \quad i = 1, \dots, I, \quad (3.2)$$

where β and x_i are as in (3.1) and

$$\{F_i/(1 - F_i)\}^{(\lambda)} = \begin{cases} \log\{F_i/(1 - F_i)\} & \text{if } \lambda = 0 \\ \lambda^{-1} [\{F_i/(1 - F_i)\}^\lambda - 1] & \text{if } \lambda \neq 0. \end{cases}$$

The model (3.2) includes as a special case ($\lambda = 0$) the logistic regression model (3.1). Guerrero and Johnson (1982) applied this model to data from the National Survey of Household Income and Expenditures in Mexico to explain the variation in female participation in the Mexican labour force. They found that a value of $\lambda = -6.63$ provided a significantly better fit than the logit model ($\lambda = 0$), the values of the standard chi-squared statistic being 4.8 (7 d.f.) and 12.8 (8 d.f.) respectively. However, they applied standard methods for binomial proportions, ignoring the survey design.

Pseudo MLE

In this section, the methods of Roberts, Rao and Kumar (1987) for the logistic regression model are extended to the power transformation model (3.2). Due to difficulties in obtaining appropriate likelihood functions for general sample designs, we use "pseudo" maximum likelihood estimates, $\hat{\beta}$ and $\hat{\lambda}$, obtained from the product binomial likelihood equations for β and λ by replacing the simple response proportion r_i/n_i with the corresponding survey estimate \hat{P}_i of P_i , and n_i/n with the corresponding survey estimate \hat{W}_i of the domain proportion W_i . Here r_i is the number of "successes" in a sample of size n_i from the i -th cell, and $n = \sum n_i$. See Guerrero and Johnson (1982), for the product binomial likelihood equations. The pseudo maximum likelihood estimates (m.l.e.), $\hat{\theta}' = (\hat{\beta}', \hat{\lambda})$, can be obtained iteratively by a quasi-Newton procedure, as in Guerrero and Johnson (1982). The fitted response proportions are given by $\hat{F} = F_i(\hat{\theta})$.

Let \hat{V}_p be the estimated covariance matrix of the survey estimates $\hat{P} = (\hat{P}_1, \dots, \hat{P}_I)'$, and let

$$B = D(\hat{F})^{-1} D(1 - \hat{F})^{-1} (\partial F / \partial \hat{\theta})'. \quad (3.3)$$

Here $D(\hat{F}) = \text{diag}(\hat{F}_i, i = 1, \dots, I)$, $D(1 - \hat{F}) = \text{diag}(1 - \hat{F}_i, i = 1, \dots, I)$ and $(\partial F / \partial \hat{\theta})'$ is the $I \times (s + 1)$ matrix of partial derivatives $\partial F_i / \partial \beta_j$ and $\partial F_i / \partial \lambda$ evaluated at $\hat{\theta}$:

$$\begin{aligned} \partial F_i / \partial \beta_j &= x_{ji} F_i^2 (1/Q_i)^{1+1/\lambda} \\ \partial F_i / \partial \lambda &= F_i^2 (Q_i \log Q_i - Q_i + 1) \lambda^{-2} (1/Q_i)^{1+1/\lambda}, \end{aligned} \quad (3.4)$$

where $Q_i = 1 + \lambda \sum_j x_{ji} \beta_j$. The estimated asymptotic covariance matrix of $\hat{\theta}$, taking account of the survey design, is then given by (see Roberts 1985)

$$\text{est cov}(\hat{\theta}) = (B' \hat{\Delta} B)^{-1} (B' D(\hat{W}) \hat{V}_p D(\hat{W}) B) (B' \hat{\Delta} B)^{-1}, \quad (3.5)$$

where $\hat{\Delta} = \text{diag}(\hat{W}_i \hat{F}_i (1 - \hat{F}_i); i = 1, \dots, I)$ and $D(\hat{W}) = \text{diag}(\hat{W}_i, i = 1, \dots, I)$.

It is also of interest to find the standard errors of the residuals $\hat{R}_i = \hat{P}_i - \hat{F}_i$ since the standardized residuals $\hat{R}_i/\text{s.e.}(\hat{R}_i)$ can be used to detect any outlying cell proportions. The estimated asymptotic covariance matrix of the vector of residuals $\hat{R} = (\hat{R}_1, \dots, \hat{R}_I)'$ is given by

$$\text{est cov}(\hat{R}) = A \text{ est cov}(\hat{\theta}) A' = \hat{V}_R, \quad (3.6)$$

where

$$A = I - D(\hat{F})D(1 - \hat{F})B(B'\hat{\Delta}B)^{-1}B'D(\hat{W}).$$

The square root of the diagonal elements, $\hat{V}_{ii,R}$, of (3.6) provide the estimated standard errors of the $\hat{R}_i, i = 1, \dots, I$.

Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of goodness-of-fit of the model (3.2) are given by

$$X^2 = n \sum_{i=1}^I (\hat{P}_i - \hat{F}_i)^2 \hat{W}_i / \{\hat{F}_i(1 - \hat{F}_i)\} \quad (3.7)$$

and

$$G^2 = 2n \sum_{i=1}^I \hat{W}_i [\hat{P}_i \log(\hat{P}_i/\hat{F}_i) + (1 - \hat{P}_i) \log\{(1 - \hat{P}_i)/(1 - \hat{F}_i)\}], \quad (3.8)$$

respectively, where the term in [] of (3.8) equals $-\log(1 - \hat{F}_i)$ at $\hat{P}_i = 0$ and $-\log \hat{F}_i$ at $\hat{P}_i = 1$.

Under product binomial sampling, it is well-known that both X^2 and G^2 are asymptotically identically distributed as a χ^2 variable with $I - s - 1$ d.f., but for general sample designs this result is no longer valid. In fact, X^2 (or G^2) is asymptotically distributed as a weighted sum, $\sum \delta_k W_k$, of independent χ^2 variables, W_k , each with 1 d.f., where the weights δ_k ($k = 1, \dots, I - s - 1$) can be interpreted as "generalized design effects" (see Roberts 1985). Under product binomial sampling, $\delta_k = 1$ for all k , and $\sum \delta_k W_k$ reduces to χ^2 with $I - s - 1$ d.f.

A first-order correction to X^2 (or G^2) is obtained by treating $X_c^2 = X^2/\hat{\delta}$ or $G_c^2 = G^2/\hat{\delta}$ as χ^2 with $I - s - 1$ d.f., where

$$(I - s - 1)\hat{\delta} = \sum \delta_k = n \sum_{i=1}^I \hat{V}_{ii,R} \hat{W}_i / \{\hat{F}_i(1 - \hat{F}_i)\} \quad (3.9)$$

and $\hat{V}_{ii,R}$ is the estimated variance of the i -th residual \hat{R}_i .

A more accurate, second order correction to X^2 (or G^2), based on the Satterthwaite approximation to $\sum \delta_k W_k$, is obtained by treating

$$X_S^2 = \frac{X_c^2}{1 + \hat{a}^2} \text{ or } G_S^2 = \frac{G_c^2}{1 + \hat{a}^2} \text{ as } \chi^2 \text{ with } (I - s - 1)/(1 + \hat{a}^2) \text{ d.f.} \quad (3.10)$$

Here $\hat{a}^2 = \sum (\hat{\delta}_k - \hat{\delta}_.)^2 / \{ (I - s - 1) \hat{\delta}_.^2 \}$ is the squared coefficient of variation of the $\hat{\delta}_i$ which can be computed, without evaluating the individual weights $\hat{\delta}_i$, from (3.9) and from

$$\sum \delta_k^2 = \sum_{i=1}^I \sum_{l=1}^I \hat{V}_{il,R}^2 (n\hat{W}_i) (n\hat{W}_l) / \{ \hat{f}_i \hat{f}_l (1 - \hat{f}_i) (1 - \hat{f}_l) \}, \tag{3.11}$$

where $\hat{V}_{il,R}$ is the (i,l) -th element of \hat{V}_R given by (3.6).

Nested hypotheses, given the model (3.2), can also be tested by correcting the standard tests for nested hypotheses, but we omit this topic for simplicity (see Roberts 1985 and Kumar and Rao 1985 for details). It is simpler, however, to use Wald tests based on the estimates $\hat{\beta}$ and the associated estimated asymptotic covariance matrix.

Example

The previous method was applied to data from the monthly Canadian Labour Force Survey (October, 1980). The Labour Force Survey design employs multi-stage cluster sampling with two stages in the self-representing urban areas and three or four stages in the non-self-representing areas in each province. A detailed description of the sample design and associated estimation procedures for the Labour Force Survey is given in Statistics Canada (1977).

The sample from the Labour Force Survey, for the present example, consisted of males aged 15-64 who were in the labour force and not full-time students. Two factors, age and education, were chosen to explain the unemployment rates via a Box-Cox transformation model. Age-group levels were formed by dividing the interval [15,64] into ten groups with the j -th age group being the interval $[10 + 5j, 14 + 5j]$ for $j = 1, \dots, 10$ and then using the mid-point of each interval, $A_j = 12 + 5j$, as the value of age for all persons in that age group. Similarly, the levels of education, E_k , were formed by assigning to each person a value based on the median years of school resulting in the following six levels: 7, 10, 12, 13, 14 and 16. The resultant age by education cross-classification provides a two-way table of $I = 60$ survey estimates, \hat{P}_{jk} , of employment rates P_{jk} . The estimated covariance matrix \hat{V}_P was based on more than 450 sample clusters.

We considered the following transformation model for $P_{jk} = F_{jk}(\theta)$ involving linear and quadratic age effects and linear education effect:

$$\begin{aligned} v_{jk}(\lambda) &= \{ F_{jk} / (1 - F_{jk}) \}^{(\lambda)} \\ &= \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \quad j = 1, \dots, 10, \quad k = 1, \dots, 6. \end{aligned} \tag{3.12}$$

Table 2 contains the pseudo m.l.e. of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \lambda)'$ and associated standard errors, and the test statistics X^2 , G^2 , X^2_S and G^2_S for testing the goodness-of-fit of the model (3.12). The corresponding values under the logistic regression model ($\lambda = 0$) are also given for comparison.

It is clear from Table 2 that the value of X^2 (or G^2) is essentially equal to the corresponding value under the logistic regression model. Thus in the present example the transformation model provides no improvement in the fit over the logistic regression model. This is also clear from the value of $\hat{\lambda}$ ($= 0.016$) which is not significantly different from $\lambda = 0$ when compared to its standard error ($= 0.085$). The estimates of regression coefficients are essentially equal under the two models, but the standard errors of the $\hat{\beta}_i$ under the Box-Cox model are much larger than the corresponding standard errors under the logistic regression model, due to the large standard error associated with $\hat{\lambda}$ and the fact that the $\hat{\beta}_i$ depend on $\hat{\lambda}$.

Table 2
Pseudo MLE of the Parameters ($\hat{\beta}', \lambda$), their Standard Errors and
Test Statistics Under the Transformation Model and under
the Corresponding Logistic Regression Model ($\lambda = 0$)

	Transformation Model		Logistic Regression Model	
	estimate	s.e.	estimate	s.e.
$\hat{\beta}_0$	- 3.28	0.975	- 3.10	0.247
$\hat{\beta}_1$	0.219	0.0468	0.211	0.013
$\hat{\beta}_2$	- 0.00227	0.00049	- 0.00218	0.00017
$\hat{\beta}_3$	0.1579	0.0385	0.1509	0.0115
$\hat{\lambda}$	0.016	0.085	—	—
Test Statistics				
	value	d.f.	value	d.f.
X^2	99.6	55	99.8	56
G_2	102.6	56	102.5	56
X^2_S	40.7	39.2	23.4	24.2
G^2_S	42.0	39.2	23.9	24.2
$X^2_S(0.05)$	54.6	55	47.7	56
$G^2_S(0.05)$	56.4	55	48.9	56

If the survey design is ignored and the value of X^2 (or G^2) is referred to $\chi^2_{0.05}(55) = 73.3$, the upper 5% point of χ^2 with $I - s - 1 = 55$ d.f., we would reject the model (3.12). On the other hand, the value of X^2_S (or G^2_S) when adjusted to refer to $\chi^2_{0.05}(55)$, denoted as $X^2_S(0.05)$ (or $G^2_S(0.05)$) in Table 2, is not significant at the 5% level, indicating that the model provides a good fit to the data, \hat{P}_{jk} .

Box and Cox (1982) and Hinkley and Runger (1984) argued that statistical inference about β should proceed with the scale determined by the estimate $\hat{\lambda}$ regarded as fixed. Thus, the estimated covariance matrix of $\hat{\beta}$ is determined from (3.5) by replacing $\partial F/\partial \hat{\theta}$ by $\partial F/\partial \hat{\beta}$ in the expression for B (equation (3.3)). For our example, this argument would suggest that we can take $\hat{\lambda} = 0$ and use the estimates of β and associated standard errors (or estimated covariance matrix) under the logistic regression model, given in Table 2.

4. TESTING EQUALITY OF LOGISTIC REGRESSION MODELS

Structural changes between two time periods may be detected through tests of equality of parameters in the corresponding models. Such tests for standard linear regression models have been developed extensively in the econometric literature (see e.g., Amemiya 1985, Sec. 1.5.3). In this section, corrected chi-squared and likelihood ratio tests of equality of parameters in two logistic regression models, corresponding to two specified time periods, are obtained. If the hypothesis of equality is tenable, then “smoothed” (i.e., fitted) estimates of cell proportions for the current period can be obtained by combining the data for the two periods.

These estimates are more efficient than the corresponding smoothed estimate based only on the current period data. The methodology is applied to data from the October 1980 and October 1981 Canadian Labour Force Survey, to study year-to-year structural changes. Note that the data for October 1980 has already been used, in Section 3, to illustrate the fitting of Box-Cox power transformation models, and it was found that a logistic regression model involving linear and quadratic age effects and linear education effect provides a good fit to the data.

Let P_{ti} be the population response proportion in the i -th cell for the period $t(=1,2)$. Then a logistic regression model for the proportions $P_{ti} = F_i(\beta_t) = F_{ti}$ is given by

$$\log\{F_{ti}/(1 - F_{ti})\} = x_i'\beta_t, \quad i = 1, \dots, I; t = 1,2 \tag{4.1}$$

where x_i is an s -vector of known constants derived from the factor levels, as in (3.1), and β_t is an s -vector of unknown parameters for period t . We are interested in testing the composite hypothesis $\beta_1 = \beta_2 (= \beta)$ to study structural changes between the two time periods. If the hypothesis is accepted, “smoothed” estimates of the proportions P_{2i} for the current period ($t = 2$) can be obtained as $F_i(\hat{\beta})$ where $\hat{\beta}$ is the pseudo m.l.e. of the common parameter β .

Pseudo MLE

Let \hat{P}_{1i} and \hat{P}_{2i} ($i = 1, \dots, I$) be the survey estimates based on sample sizes n_1 and n_2 respectively. Extending the notation in Section 3, “pseudo” maximum likelihood estimates, $\hat{\beta}_t$, are obtained from the product binomial likelihood equations for β_t by replacing the simple response proportions r_{ti}/n_{ti} with the corresponding survey estimates \hat{P}_{ti} of P_{ti} and n_{ti}/n_t with the corresponding survey estimates \hat{W}_{ti} of the domain proportions W_{ti} , thus yielding

$$X'D(\hat{W}_t)\hat{F}_t = X'D(\hat{W}_t)\hat{P}_t, \quad t = 1,2 \tag{4.2}$$

where $\hat{F}_t = F(\hat{\beta}_t)$ is the vector of fitted response proportions for period t , $D(\hat{W}_t) = \text{diag}(\hat{W}_{ti}, i = 1, \dots, I)$, and $X' = (x_1, \dots, x_I)$. The estimates $\hat{\beta}_t$ are obtained iteratively by a quasi-Newton procedure.

Under the hypothesis $\beta_1 = \beta_2 (= \beta)$, the pseudo maximum likelihood estimates, $\hat{\beta}$, are obtained by iteration from the following pseudo likelihood equations:

$$X'D(\hat{W}_c)\hat{\hat{F}} = (n_1/n)X'D(\hat{W}_1)\hat{P}_1 + (n_2/n)X'D(\hat{W}_2)\hat{P}_2, \tag{4.3}$$

where $D(\hat{W}_c) = (n_1/n)D(\hat{W}_1) + (n_2/n)D(\hat{W}_2)$, $\hat{\hat{F}} = F(\hat{\hat{\beta}})$ is the vector of fitted response proportions or smoothed estimates of cell proportions for the current period, and $n_1 + n_2 = n$.

Let \hat{V}_P be the estimated covariance matrix of $(\hat{P}_1', \hat{P}_2')'$ partitioned as

$$\hat{V}_P = \begin{bmatrix} \hat{V}_{11P} & \hat{V}_{12P} \\ \hat{V}_{21P} & \hat{V}_{22P} \end{bmatrix}.$$

Then the estimated covariance matrix of smoothed estimates $\hat{\hat{F}}$ is given by

$$\text{est cov}(\hat{\hat{F}}) = B\hat{V}_PB', \tag{4.4}$$

where

$$B = D(\hat{W}_c)^{-1} \hat{\Delta} X (X' \hat{\Delta} X)^{-1} X' [(n_1/n) D(\hat{W}_1), (n_2/n) D(\hat{W}_2)] \quad (4.5)$$

and

$$\hat{\Delta} = \text{diag}(\hat{W}_c \hat{F}_i (1 - \hat{F}_i)), i = 1, \dots, I.$$

If the residuals are defined as $\hat{R}_t = \hat{F}_t - \hat{F}$, then the estimated covariance matrix of $(\hat{R}'_1, \hat{R}'_2)'$ is given by

$$\hat{V}_R = \begin{bmatrix} \hat{V}_{11R} & \hat{V}_{12R} \\ \hat{V}_{21R} & \hat{V}_{22R} \end{bmatrix} = A \hat{V}_P A'. \quad (4.6)$$

Here

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with

$$A_{t1} = D(\hat{W}_c)^{-1} \hat{\Delta} X \left[(X' \hat{\Delta}_t X)^{-1} X' D(\hat{W}_t) - \frac{n_t}{n} (X' \hat{\Delta} X)^{-1} X' D(\hat{W}_t) \right],$$

and

$$A_{t2} = -D(\hat{W}_c)^{-1} \hat{\Delta} X (X' \hat{\Delta} X)^{-1} X' \left\{ D(\hat{W}) - \frac{n_t}{n} D(\hat{W}_t) \right\}, t = 1, 2,$$

where

$$\hat{\Delta}_t = \text{diag}(\hat{W}_{ti} \hat{F}_i (1 - \hat{F}_i)), i = 1, \dots, I.$$

Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of the nested hypothesis $\beta_1 = \beta_2$, given the model (4.1), are given by

$$X^2 = X_1^2 + X_2^2 \quad (4.8)$$

and

$$G^2 = G_1^2 + G_2^2, \quad (4.9)$$

where

$$X_t^2 = n_t \sum_{i=1}^I (\hat{F}_{ti} - \hat{F}_i)^2 \hat{W}_{ti} / \{\hat{F}_i (1 - \hat{F}_i)\}, t = 1, 2 \quad (4.10)$$

and

$$G_t^2 = 2n_t \sum_{i=1}^I \hat{W}_{ti} \left[\hat{F}_{ti} \log(\hat{F}_{ti} / \hat{F}_i) + (1 - \hat{F}_{ti}) \log\{(1 - \hat{F}_{ti}) / (1 - \hat{F}_i)\} \right], t = 1, 2. \quad (4.11)$$

A first order correction to X^2 (or G^2) is obtained by treating $X_c^2 = X^2/\hat{\delta}$, or $G_c^2 = G^2/\hat{\delta}$, as χ^2 with s d.f., where

$$s\hat{\delta} = n_1 \sum_{i=1}^I \hat{V}_{11R}(ii) \hat{W}_{1i} / \{\hat{F}_i(1 - \hat{F}_i)\} + n_2 \sum_{i=1}^I \hat{V}_{22R}(ii) \hat{W}_{2i} / \{\hat{F}_i(1 - \hat{F}_i)\} \quad (4.12)$$

and $\hat{V}_{iIR}(ij)$ is the (i,j) th element of \hat{V}_{iIR} . A more accurate, second order correction to X^2 (or G^2), based on the Satterthwaite approximation, is obtained by treating

$$X_S^2 = \frac{X_c^2}{1 + \hat{a}^2} \quad \text{or} \quad G_S^2 = \frac{G_c^2}{1 + \hat{a}^2} \quad \text{as} \quad \chi^2 \quad \text{with} \quad s/(1 + \hat{a}^2) \quad \text{d.f.} \quad (4.13)$$

Here $\hat{a}^2 = (\sum_{k=1}^s \hat{\delta}_k^2 - s\hat{\delta}^2)/s\hat{\delta}^2$ which can be computed from (4.12) and the following formula for $\sum \hat{\delta}_k^2$:

$$\begin{aligned} \sum_{k=1}^s \hat{\delta}_k^2 = & n_1^2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{11R}^2(ij) \hat{W}_{1i} \hat{W}_{1j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)} \\ & + n_2^2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{22R}^2(ij) \hat{W}_{2i} \hat{W}_{2j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)} \\ & + 2n_1 n_2 \sum_{i=1}^I \sum_{j=1}^I \frac{\hat{V}_{12R}^2(ij) \hat{W}_{1i} \hat{W}_{2j}}{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)}, \end{aligned} \quad (4.14)$$

where $\hat{V}_{12R}(ij)$ is the (i,j) -th element of \hat{V}_{12R} .

Example

The previous method was applied to data from the October 1980 and October 1981 Canadian Labour Survey, to study year-to-year structural changes.

The logistic regression model involving linear and quadratic age effects and linear education effect provided a good fit to data from both periods with the following estimates of β_t :

$$\hat{\beta}_1: \{-3.08, 0.211, -0.00218, 0.1505\}$$

$$\hat{\beta}_2: \{-3.05, 0.179, -0.00169, 0.1707\},$$

where $\log\{\hat{F}_{ijk}/(1 - \hat{F}_{ijk})\} = \hat{\beta}_{t0} + \hat{\beta}_{t1}A_j + \hat{\beta}_{t2}A_j^2 + \hat{\beta}_{t3}E_k, j = 1, \dots, 10; k = 1, \dots, 6$ and \hat{F}_{ijk} is the fitted employment rate in the (j,k) -th cell for period t . One cell was omitted in the fitting since the domain sample size n_{2t} is zero for the current period.

Turning to the test of the hypothesis $\beta_1 = \beta_2$, given the logistic regression models, we obtained the following values of X^2, G^2, X_c^2, G_c^2 and X_S^2, G_S^2 :

$$X^2 = 42.1 \quad X_c^2 = 24.6 \quad X_S^2 = 24.4$$

$$G^2 = 42.2 \quad G_c^2 = 24.6 \quad G_S^2 = 24.4.$$

Also $s/(1 + \hat{a}^2) = 4/(1.0089) = 3.965 \doteq 4$. By referring X_S^2 or G_S^2 to $\chi_{0.05}^2(4) = 9.49$, the upper 5% point of χ^2 with 4 d.f., we reject the hypothesis $\beta_1 = \beta_2$ at the 5% level, indicating significant year-to-year structural changes for the month of October. The data for the two time periods, therefore, should not be pooled to get smoothed estimates of unemployment rates, $1 - \hat{F}_{jk}$, for the current period.

5. POLYTOMOUS RESPONSE MODELS

A variety of models has been suggested in the literature when the response variable is polytomous. The variety of models reflects, in part, the different scales of measurement possible for polytomous response variables, unlike binary response variables. In the main, there are nominal responses where any permutation of the response categories is equally valid, and ordinal responses where there is a natural ordering of the response categories.

Suppose that the population of interest is partitioned into I cells (or domains) according to the levels of one or more factors. Let $P_{j(i)}$ be the population proportion in the i^{th} cell having the j^{th} response ($j = 1, \dots, J+1$) so that $\sum_{j=1}^{J+1} P_{j(i)} = 1$ ($i = 1, \dots, I$). Then a general polytomous response model for the proportions $P_j(i)$ is given by

$$P_j(i) = F_{ij}(\theta), \quad i = 1, \dots, I; \quad j = 1, \dots, J, \quad (5.1)$$

where θ is an r -vector of unknown parameters ($r \leq IJ$) and $F_{ij}(\theta)$ is a function of known form. In the nominal case, Haberman (1982) and others proposed the following model: the "multinomial logits" $\log P_j(i) - \sum_{j'=1}^{J+1} \log P_{j'(i)} (J+1)^{-1}$ are assumed to be unknown linear functions of x_i , the s -vector of known constants derived from the factor levels, *i.e.*,

$$F_{ij}(\theta) = \exp(x_i' \beta_j) / \sum_{k=1}^{J+1} \exp(x_i' \beta_k), \quad i = 1, \dots, I; \quad j = 1, \dots, J+1 \quad (5.2)$$

with $\sum \beta_k = 0$. Because of the latter constraint on the β_k , (5.2) may be expressed as

$$F_{ij}(\theta) = \exp(x_i' \beta_j) / \left[\sum_{k=1}^J \exp(x_i' \beta_k) + \prod_{k=1}^J \exp(-x_i' \beta_k) \right],$$

$$i = 1, \dots, I; \quad j = 1, \dots, J. \quad (5.3)$$

Note that (5.3) reduces to the usual logistic regression model in the special case of binary response.

In the ordinal case, a simple model which also has the feature of being invariant under the grouping of response categories is given by (McCullagh 1980)

$$\log\{C_{j(i)}/(1 - C_{j(i)})\} = \nu_j - x_i' \beta, \quad j = 1, \dots, J; \quad i = 1, \dots, I \quad (5.4)$$

where $C_{j(i)} = \sum_{k=1}^j P_{k(i)}$ denotes the j^{th} cumulative probability in the i^{th} domain, and $\theta' = (\nu_1, \dots, \nu_J, \beta')$. To express (5.4) in the form (5.1), we note that $P_i = L^{-1}C_i$, where $P_i = (P_{1(i)}, \dots, P_{J(i)})'$, $C_i = (C_{1(i)}, \dots, C_{J(i)})'$ and L^{-1} is a $J \times J$ nonsingular matrix with 1 in the diagonal, -1 in the $(i+1, i)^{\text{th}}$ position ($i < J$) and 0 elsewhere.

Pseudo MLE

As before, we use pseudo m.l.e., $\hat{\theta}$ obtained from the product multinomial likelihood equations for θ by replacing the simple response proportions n_{ij}/n_i with the corresponding survey estimates $\hat{P}_{j(i)}$, and n_i/n with the corresponding survey estimate \hat{W}_i of the domain proportion W_i . Here n_{ij} is the number of units with the j^{th} response in a sample of size n_i from the i^{th} domain and $n = \sum n_i$. The fitted response proportions are then given by $\hat{F} = F(\hat{\theta}) = (\hat{F}'_1, \dots, \hat{F}'_I)'$, where $\hat{F}_i = (\hat{F}_{i1}, \dots, \hat{F}_{iJ})'$ and $\hat{F}_{ij} = F_{ij}(\hat{\theta})$.

Let \hat{V}_P be the estimated covariance matrix of the survey estimates $\hat{P} = (\hat{P}_{1(1)}, \dots, \hat{P}_{J(1)}, \dots, \hat{P}_{1(I)}, \dots, \hat{P}_{J(I)})'$, and $\hat{M} = (\partial F/\partial \hat{\theta})'$, the $IJ \times r$ matrix of partial derivatives $\partial F_{ij}/\partial \theta_k$ calculated at $\hat{\theta}$. Also, let $\hat{Q}_i = \text{diag}(\hat{F}_i) - \hat{F}_i \hat{F}_i'$ and $\hat{Q} = \text{diag}(\hat{Q}_i, i = 1, \dots, I)$. The expressions for the partial derivatives $\partial F_{ij}/\partial \theta_k$ for the models (5.3) and (5.4) are given in Roberts (1985). The estimated asymptotic covariance matrix of $\hat{\theta}$, taking account of the survey design, is then given by (see Roberts 1985).

$$\text{est cov}(\hat{\theta}) = (\hat{M}' \hat{\nabla} \hat{M})^{-1} (\hat{M}' \hat{\nabla} \hat{V}_P \hat{\nabla}' \hat{M}) (\hat{M}' \hat{\nabla} \hat{M})^{-1}, \tag{5.5}$$

where $\hat{\nabla} = (D(\hat{W}) \otimes I) \hat{Q}^{-1}$ and $D(\hat{W}) = \text{diag}(\hat{W}_i, i = 1, \dots, I)$. In the special case of product multinomial sampling, $\hat{V}_P = \hat{\nabla}^{-1}/n$ and (5.5) reduces to $(\hat{M}' \hat{\nabla} \hat{M})^{-1}/n$.

The vector of residuals, $\hat{R} = \hat{P} - \hat{F}$, is also of interest, since it may be useful in detecting model deviations. The estimated asymptotic covariance matrix of \hat{R} is given by

$$\text{est cov}(\hat{R}) = \hat{G} \hat{V}_P \hat{G}' \tag{5.6}$$

where $\hat{G} = I - \hat{M}(\hat{M}' \hat{\nabla} \hat{M})^{-1} \hat{M}' \hat{\nabla}$.

Corrections to standard tests

For simplicity, we consider only the Pearson chi-squared test of goodness-of-fit of the model (5.1). It is given by

$$X^2 = n \sum_{i=1}^I \hat{W}_i \sum_{j=1}^{J+1} (\hat{P}_{j(i)} - \hat{F}_{ij})^2 / \hat{F}_{ij}. \tag{5.7}$$

Under independent multinomial sampling in each of the domains, it is well-known that X^2 is asymptotically distributed as a χ^2 variable with $IJ - r$ d.f.

To test the nested hypothesis $\theta_2 = 0$, given the model (5.1), let $\hat{\theta}_1$ be the pseudo m.l.e. of θ_1 and \hat{F} be the corresponding vector of fitted response proportions, where $\theta' = (\theta'_1, \theta'_2)$, θ_1 is $q \times 1$ and θ_2 is $u \times 1$ ($q + u = r$). The Pearson chi-squared test of the nested hypothesis is then given by

$$X^2(2|1) = n \sum_{i=1}^I \hat{W}_i \sum_{j=1}^{J+1} (\hat{F}_{ij} - \hat{\hat{F}}_{ij})^2 / \hat{\hat{F}}_{ij} \tag{5.8}$$

which is asymptotically distributed as χ^2 with u d.f. under independent multinomial sampling in each of the domains. However, for a general sample design, X^2 and $X^2(2|1)$ are both asymptotically distributed as weighted sums of independent χ^2 variables, each with 1 d.f., where the weights can be interpreted as “generalized design effects” of particular linear transformations of \hat{P} (Roberts 1985).

A first-order correction to $X^2(2|1)$ is obtained by treating

$$X_c^2(2|1) = X^2(2|1)/\hat{\delta} \cdot (2|1) \text{ as } \chi^2 \text{ with } u \text{ d.f.,} \tag{5.9}$$

where $\hat{\delta} \cdot (2|1)$ is obtained by replacing θ' by $(\hat{\theta}'_1, 0')$ and V_P by \hat{V}_P in the following definition for $\delta \cdot (2|1)$:

$$u\delta \cdot (2|1) = \sum_{i=1}^u \delta_i(2|1) = \text{tr } D(2|1). \tag{5.10}$$

Here, tr denotes the trace operator and $D(2|1)$ is a generalized design effects matrix given by

$$D(2|1) = (H_2' \nabla H_2)^{-1} (H_2' \nabla V_P \nabla' H_2), \tag{5.11}$$

where V_P is the covariance matrix of \hat{P} , $\nabla = (D(W) \otimes I)Q^{-1}$, Q is the block diagonal matrix with $Q_i = \text{diag}(F_i) - F_i F_i'$, $i = 1, \dots, I$, $F_i = F_i(\theta)$, and $H_2 = [I - M_1 (M_1' \nabla M_1)^{-1} M_1' \nabla] M_2$, where $M_1 = (\partial F / \partial \theta_1)'$ and $M_2 = (\partial F / \partial \theta_2)'$.

A more accurate, second order correction to $X^2(2|1)$, based on the Satterthwaite approximation, is obtained by treating

$$X_s^2(2|1) = X_c^2(2|1) / [1 + \hat{a}(2|1)^2] \text{ as } \chi^2 \text{ with } u / [1 + \hat{a}(2|1)^2] \text{ d.f.} \tag{5.12}$$

Here $\hat{a}(2|1)^2$ is obtained by replacing θ by $(\hat{\theta}'_1, 0')$ in the following definition of $a(2|1)^2$:

$$a(2|1)^2 = \left\{ \sum_{i=1}^u \delta_i(2|1)^2 - u\delta \cdot (2|1)^2 \right\} / u\delta \cdot (2|1)^2, \tag{5.13}$$

where

$$\sum_{i=1}^u \delta_i(2|1)^2 = \text{tr} D(2|1)^2. \tag{5.14}$$

The corrections to goodness-of-fit test X^2 are obtained as special cases of (5.9) and (5.12) by treating the model as nested within a saturated model (*i.e.*, a model where the unknown parameter θ is of length IJ).

Example

The previous methods were applied to data from the Canada Health Survey (1978-79). A brief description of the survey is provided in Section 2.

The data set examined consisted of the estimated counts of females aged 20-64 cross-classified by frequency of breast self-examination (with the 3 categories: monthly, quarterly, less often or never), education (with the 3 categories: secondary or less, some post-secondary, post-secondary) and age (with the 3 categories: 20-24, 25-44, 45-64).

The frequency of breast self-examination was considered to be the response variable, while education and age were taken as explanatory variables, so that the number of responses, $J + 1$, equalled 3 and the number of domains, I , was 9. Both response and explanatory variables are ordered.

Table 3
Survey Estimates of Cumulated Probabilities

	Age	Education	$C_{1(ik)}$	$C_{2(ik)}$
$i = 1, k = 1$	20-24	\leq Secondary	.25	.49
$k = 2$		$<$ Post-Secondary	.25	.41
$k = 3$		\geq Post-Secondary	.23	.47
$i = 2, k = 1$	25-44	\leq Secondary	.25	.50
$k = 2$		$<$ Post-Secondary	.27	.44
$k = 3$		\geq Post-Secondary	.26	.44
$i = 3, k = 1$	45-64	\leq Secondary	.28	.51
$k = 2$		$<$ Post-Secondary	.24	.62
$k = 3$		\geq Post-Secondary	.29	.56

Table 4
Statistics for Testing Goodness of Fit and Nested Hypotheses

	Goodness of Fit (Age & Education)	Nested Hypothesis (Age only)
X^2	37.7	7.1
X_c^2	21.6	3.8
X_S^2	18.5*	3.7*
$\hat{\delta}$	1.75	1.9
\hat{a}^2	0.83	0.1

* The Satterthwaite statistic has been adjusted to refer to the same χ^2 value as X_c^2 .

The following model for the cumulated probabilities of the type described in equation (5.4), was considered:

$$\log\{C_j(ik)/(1 - C_j(ik))\} = \nu_j + \beta a_i + e_k \quad (j = 1,2; i = 1,2,3; k = 1,2,3) \quad (5.15)$$

where $C_j(ik)$ is the j^{th} cumulated probability for the i^{th} age group and k^{th} education group. As well, $a_i = A_i - \bar{A}$, where A_i is the midpoint of the i^{th} age interval, and e_k is the effect of the k^{th} education group ($\sum e_k = 0$), ignoring the order of the education categories. Table 3 contains the survey estimates of the cumulated proportions. Table 4 contains the test statistics X^2 , X_c^2 and X_S^2 for testing the goodness of fit of (5.15) and also for testing the nested hypothesis of no education effect, $e_k = 0$ for $k = 1,2$.

First, considering the goodness of fit of (5.15), if the survey design is ignored and the value of X^2 is referred to $\chi^2_{0.05}(13) = 22.4$, the upper 5% point of χ^2 with $IJ - 5 = 13$ d.f., we would reject the model. On the other hand, the value of X_c^2 or the value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(13)$, is not significant at the 5% level, indicating that the model provides a good fit to the data.

For testing of the nested hypothesis, the value of X_c^2 , or the value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(2) = 5.99$ is not significant at the 5% level, indicating that the nested hypothesis of no education effect is tenable.

6. SOFTWARE

Implementation of the methodology of the previous sections requires two stages of computation — calculation of a vector of proportions, along with its estimated covariance matrix, and then calculation of model estimates, test statistics and their adjustments.

Surveys like the Canada Health Survey and the Labour Force Survey, from which examples have been presented, have complex designs and large data bases. Because of these two factors, calculation of covariance matrices was done on a mainframe computer. Custom SAS and Fortran programs were used for this purpose.

Computations required for the fitting and testing of goodness-of-fit models and sub-hypotheses were done either on the mainframe computer using SAS (and the MATRIX procedure in particular), or on a microcomputer using the GAUSS programming package.

These programs are available to other analysts at Statistics Canada.

REFERENCES

- AMEMIYA, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.
- BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, 279-292.
- BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- BOX, G.E.P., and COX, D.R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209-210.
- FAY, R.E. (1985). A jack-knifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.
- FULLER, W.A. (1986). Estimators of the factor model for survey data. In *Advances in the Statistical Sciences*, Vol. I (Eds. MacNeill, I.B. and Umphrey, G.J.). Dordrecht, Holland: Reidel Publishing Co., 265-284.
- GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series B*, 46, 270-272.
- GUERRERO, V.M., and JOHNSON, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, 69, 309-314.
- HABERMAN, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.
- HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys, Parts I and II. *Journal of Official Statistics*, 3, 117-132 and 133-140.
- HINKLEY, D.V., and RUNGER, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302-309.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- KOCH, G.G., FREEMAN, D.H. Jr., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KUMAR, S., and RAO, J.N.K. (1985). Fitting Box-Cox transformation models to labour force survey data. Unpublished Report, Social Surveys Methods Division, Statistics Canada, Ottawa.

- McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109-142.
- NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the American Statistical Association*, 76, 681-689.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- ROBERTS, G. (1985). *Contributions to Chi-Squared Tests with Survey Data*. Unpublished Ph.D. Thesis, Carleton University, Department of Mathematics and Statistics, Ottawa.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SCOTT, A.J. (1986). Logistic regression analysis with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 25-30.
- SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- SCOTT, A.J., RAO, J.N.K., and THOMAS, D.R. (1989). Weighted least squares and quasi maximum likelihood estimation for categorical data under generalized linear models. *Linear Algebra and its Applications*, second special issue on Linear Algebra and Statistics, in press.
- SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Statistics Canada Working Paper No. SSMD 86-002.
- SINGH, A.C., and KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-257.
- SKINNER, C.J., HOLMES, D.J., and SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.
- STATISTICS CANADA (1977). *Methodology of the Canadian Labour Force Survey, 1976*. Catalogue 71-526 occasional. Ottawa: Statistics Canada.
- THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

COMMENT

ROBERT E. FAY¹

The authors have made an excellent contribution to the literature on the analysis of data from complex samples. By examining in turn four different models for categorical data: i) a log-linear model for a cross-classification; ii) a modification of the approach of Box and Cox to the transformation of binary data; iii) a problem of inference about parameters of a logistic regression model; and iv) a polytomous response model, the authors present solutions to important individual problems and illustrate the ways in which these flexible approaches to inference can be extended to other models for categorical data from complex samples. The applications are connected by an underlying theory, much of it previously appearing in Rao and Scott (1984), but this paper usefully presents in greater detail the implications of the general theory for specific models.

An omission from the paper is understandable but worth noting: for each model illustrated in the paper, replication provides an alternative strategy that, at times, may also be more convenient. In particular, the replication theory is complete for each of the applications, i), ii), and iv), to cross-classified data. In each case, tests of overall fit and comparisons of nested models can be assessed with the jackknifed chi-square test (Fay 1985) and standard errors for the parameters obtained through replication.

Replication also can provide standard errors and covariances for parameters of logistic regression models, as in iii), enabling in some cases a Wald-type test for equality of sets of regression parameters. It also appears likely that the jackknifing approach extends to the likelihood-ratio chi-square test in such situations involving continuous variables, although a firm proof of this conjecture is clearly required before application can be recommended. My point in calling attention to replication as a competing strategy for the problems presented in the paper is not to imply that it represents a methodologically superior approach to the methods of Rao and Scott (1984); instead, the availability of this methodology provides an additional choice to solve these and similar problems of inference. For example, the focus on replication for the estimation of variances from the current demographic surveys at the U.S. Census Bureau provides the potential to carry out analyses such as those presented in the paper.

I also want to point out that the methods presented and the analogues from replication theory have a potential importance beyond the realm of design-based inference from complex sample surveys, which is the focus of the paper. One of these involves the use of multiple imputation or related approaches intended to represent the uncertainty due to missing data. The implied interpretation of variance within the domain of design-based inference can be extended to include uncertainty from missing data without requiring changes to the methodology presented in the paper. The general methodology may also be applicable to some problems of inference from complex designed experiments, in which the design poses problems of clustering or stratification similar to complex sample surveys.

Of the four models discussed, however, I suggest that the Box and Cox transformation not be applied without consideration of alternative strategies, such as transformation of the x-variables instead. My own inclination would be to favor an analysis on a logistic scale, with possibly transformed predictors, unless the adaptation of the Box and Cox transformation obtains some distinct advantage, such as offering an additive model on the transformed scale in an instance where the logistic model does not provide as successful a fit without interaction terms.

I am delighted to have the opportunity to commend the authors on a useful and instructive paper.

¹ Robert E. Fay, U.S. Bureau of the Census, Washington, D.C. 20233.

COMMENT

C.J. SKINNER¹

This paper provides an excellent discussion of a variety of applications of weighted least squares (WLS) and pseudo maximum likelihood (PML) procedures to categorical data. Its clear presentation and use of real survey examples will, I hope, help to encourage survey analysts to take account of complex designs in their analyses. As the authors indicate, analytical statistical procedures which take account of complex designs have been developed extensively in recent years (see *e.g.* Skinner, Holt and Smith 1989) and are even beginning to be referred to in standard computer software (*e.g.* SAS 1985, pp 61-67).

Commenting first on some specific aspects of the paper, I found Section 5 on polytomous variables to be especially valuable, given the wide occurrence of such data in surveys. A property of ordinal variables is that they may often be expected to possess monotonic relationships and so, for example, lack of monotonicity between the fitted values of $C_{1(ik)}$ (or $C_{2(ik)}$) and the education variable k in Table 3 makes the result of the corrected tests, that there is no evidence of an education effect, more plausible than the result of the uncorrected test.

The discussion of testing equality of two logistic regression models in Section 4 also seemed to me to be practically useful, although it would still seem to be possible theoretically to formulate this test as one of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987).

Section 3 provides a useful illustration of how PML may be applied to general parametric models for categorical data. It is, however, gratifying that the more complex transformation model provides no significant improvement in fit over the logistic regression model, since the interpretation of the parameters of the transformation model is more difficult. For example, for the logistic model the coefficient for education may be interpreted as implying that the odds of being employed are increased by 16% for each additional year of education for males of a given age ($\exp(.1509) = 1.16$), whereas this interpretation is not generally available for the transformation model when $\lambda \neq 0$.

On a more general note I would be interested in the authors' views on the relative merits of WLS and PML. In the paper, these methods are presented quite separately, although both procedures would seem to be potentially applicable to a very wide class of models for categorical data under complex designs. Indeed both procedures are also applicable to models with continuous variables (Skinner, Holt and Smith 1989, Chapter 3); WLS requires just a statistic consistent for a known function of the parameters together with a consistent estimate of the covariance matrix of the statistic (Fuller 1984, Corollary 2), whereas PML is applicable very widely as described in Binder (1983). As a basis for discussion I list below a number of criteria on which WLS and PML might be compared; M1-M3 are relevant even under multinomial sampling, C1-C3 are specific to complex designs.

- M1 **Flexibility** WLS may be more adaptable than PML for complex problems *e.g.* involving structural zeros.
- M2 **Computation** WLS computation tends to have a more standard form.
- M3 **Small cell counts** WLS is more sensitive to small counts, especially zeros.
- C1 **Adaptability of multinomial methods to complex designs** WLS seems more easily adaptable.

¹ C.J. Skinner, University of Southampton, United Kingdom.

- C2 **Efficiency** Under multinominal sampling WLS is usually asymptotically equivalent to PML (which is then just standard ML). It might be conjectured that WLS will always be at least as efficient as PML under complex designs, although this presupposes a 1-1 correspondence between WLS and PML estimation problems. If WLS is more efficient, is the gain usually negligible (*cf.* Scott and Holt 1982)? Are there general results here?
- C3 **Degrees of freedom** WLS estimators and associated Wald tests may be unstable if the degrees of freedom used to estimate V_p are low.

ADDITIONAL REFERENCES

- FULLER, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* 10, 97-118.
- SAS Institute Inc. (1985). *SAS/IML User's Guide, Version 5 Edition*. Cary NC: SAS Institute Inc.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F., Eds. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

COMMENT

E.A. MOLINA¹

I would like to congratulate the authors on bringing together some recent methods developed for analyzing categorical data arising from sample surveys. The paper should be extremely useful for survey analysts who wish to take into account the impact of survey designs on the practical aspects of the analysis of survey data. In particular, it is important to emphasize that the methods discussed cover two different situations arising in practice: so called *primary analyses*, in which the researcher has all the relevant information at hand, and *secondary analyses*, in which the data provided do not include enough information about the population units to enable the calculation of full covariance matrices of the sample estimators.

The methods covered require the existence of a structural model for the data. There are situations, however, in which it is difficult to specify a single structural model that adequately describes categorical data. In large scale surveys there is often need to screen out many cross classifications at minimal cost. In such cases the use of measures of association is a common alternative. These non parametric methods were extended to sample survey data by Molina and Smith (1986, 1988).

For the primary analysis of survey data the paper concentrates on weighted least squares and Wald tests. The results in Scott, Rao and Thomas (1989) are summarized and the relationship with quasi-likelihood is mentioned. I think that an important conclusion from that paper should be included in this section, namely the need to take into account the survey constraints $K'p(X\beta) = \pi$ when using quasi-likelihood methods. The reader may not be aware of the importance of the careful choice of the g -inverse in equation (2.9). Quasi-likelihood methods are now widely used and the relationship with weighted least squares methods is a relevant one. In fact, quasi-likelihood functions represent an interesting alternative for the analysis of survey data. However, there are practical problems since the method requires that we specify the covariance matrix as a function of p , the variance function. Quasi-likelihoods are largely determined by these variance functions (see, *e.g.*, Morris 1982, and Jørgensen 1987). If a matrix of estimates is given instead of a function, the method would be equivalent to the use of a normal distribution.

Most of the paper is devoted to methods involving *pseudo likelihoods*. Since secondary analyses constitute the most common situation in practice, the methods presented are likely to be extensively used by survey analysts. I would like, however, to discuss some alternatives.

The study of the impact of survey design on Guerrero and Johnson's (1982) transformation models is an important addition to the literature. However, Nelder and Pregibon (1987) have proposed a family of functions, the *extended quasi-likelihoods*, that avoid some important disadvantages of transformation models and can be fitted with GLIM. If design effects are available, their methods can be adapted to survey data by incorporating them either in the variance functions or in the form of weights. Alternatively, design variables may be used to adjust the dispersion parameter in the models. In both cases, one advantage is that we can use the goodness of fit statistics and standard errors produced by GLIM under these models to examine the data without the introduction of further corrections.

These comments apply in general to the use of pseudo-likelihoods. The effect of ignoring the survey design may be treated as an increase or decrease in the expected variability that may be modelled as overdispersion or underdispersion by means of quasi-likelihoods or extended quasi-likelihoods. See, *e.g.*, Pocock *et al.* (1981), Breslow (1984), Williams (1982), among

¹ E.A. Molina, Universidad Simon Bolivar, Caracas and University of Southampton, United Kingdom.

others. As an example, I reanalyzed the data in Table 1. The analysis given in the paper is the correct one, since it incorporates the true covariance matrix. Suppose, however, that this matrix is not available and that only the cell design effects are at hand. Using GLIM I fitted model (2.12) with a Poisson error ignoring the sampling scheme. This gives $X^2 = 5.68$, $G^2 = 5.67$. The Rao and Scott (1987) approximation for the chi square statistic gives $X^2(\delta) = 5.68/2.25 = 2.52$. For the independence model the uncorrected values are $X^2 = 18.22$, $G^2 = 18.22$, and the correction gives $X^2(\delta) = 18.22/1.65 = 11.04$. What can be done if the deffs are not available?. A simple quasi-likelihood approach to overdispersion is to estimate the mean deviance for the larger model, $D = 5.68/3 = 1.89$, and to use the inverse of this value as a weight (or as a new scale parameter). This give $X^2 = 3.01$ for model (2.12) and $X^2 = 9.65$ for the independence model. The correct approach here is to use the excess in deviance (the difference between the log-likelihood ratio statistics) to test $\gamma = 0$, since G^2 will equate the degrees of freedom for the larger model. The value is 6.65, which is just significant at the 1% level. Both analyses are in agreement with the correct analysis given in the paper, but in other situations it may not be so. The quasi-likelihood model presented here is equivalent to assuming that the actual covariance matrix is a multiple of the one obtained under multinomial sampling, a model that may perform badly in several situations. The advantage is that it can be used when the only information available is that given by the variability inherent in the data, and the analysis performed in a standard statistical package like GLIM. If the deffs are available, other models involving them may be proposed, and a paper is in preparation.

There is, however, no completely satisfactory substitute for an analysis involving the actual covariance matrix. The objective of this contribution is to highlight other possibilities when the full covariance matrix is not known. Quasi-likelihoods offer a fertile ground for further exploration, particularly in relation to survey data. The paper under discussion presents several alternatives and is an important contribution to the field.

ADDITIONAL REFERENCES

- JØRGENSEN, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B* 127-162.
- MOLINA, E.A., and SMITH, T.M.F. (1986). The effect of sample design on the comparison of associations. *Biometrika* 73, 23-33.
- MOLINA, E.A., and SMITH, T.M.F. (1988). The effect of sampling on operative measures of association. *International Statistical Review* 56, 235-242.
- MORRIS, C.N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics* 10, 65-80.
- NELDER, J.A., and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* 74, 221-232.
- POCOCK, S.J., COOK, D.G., and BERESFORD, S.A.A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Applied Statistics* 31, 286-295.
- WILLIAMS, D.A. (1982). Extra binomial variation in logistic-linear models. *Applied Statistics* 31, 144-148.

RESPONSE FROM THE AUTHORS

We thank the three discussants, Fay, Molina and Skinner, for their useful comments and for suggesting additional methods useful in the analysis of cross-classified data from complex sample surveys.

(i) Response to comments of R.E. Fay

We agree with Fay that replication methodology and associated jackknife chi-squared tests provide viable alternatives to the methods presented here, provided the survey design permits the use of a replication method such as the jackknife or the balanced half-sample replication. His CPLX program indeed offers a comprehensive analysis option whenever estimates are available at the individual replicate level. Also, as noted in the Introduction, Fay's jackknife tests and Rao-Scott corrections have performed well under quite general conditions in simulation studies, unlike the Wald tests based on weighted least squares. Rao-Scott corrections are, however, also applicable to survey designs not permitting the use of a replication method.

The software systems for the Canada Health Survey and the Canadian Labour Force Survey were set up to readily provide the estimated covariance matrix of cell estimates but not the replicate level estimates. As a result, the implementation of jackknife tests would have required some changes in the software systems.

We are also thankful to Fay for pointing out that the methods presented here, and the analogues from replication theory, can also handle some problems of inference from complex designed experiments involving clustering and stratification. Indeed, one of us (J.N.K. Rao) recently used Rao-Scott type methods to fit dose-response models and to test hypotheses in teratological studies involving animal litters as experimental units (Rao and Colin 1989). These methods do not assume specific models for the intra-litter correlations, unlike other methods proposed in this area.

We considered Box-Cox transformation models since Guerrero and Johnson (1982) obtained significantly better fits on some Mexican data compared to the logit model. We agree with Fay, however, that the Box-Cox models should not be applied without consideration of alternative strategies, such as transforming the predictors. As noted by Fay, the Box-Cox approach would be useful in these cases where it would lead to additive models on the transformed scale while the logit model would require interaction terms.

(ii) Response to comments of E.A. Molina

Molina is correct in saying that measures of association can be used to screen out many cross classifications at minimal cost. His joint work with T.M.F. Smith on extending the classical theory for measures of association to sample survey data involving clustering and stratification is an important contribution.

As noted in the Introduction, we assumed throughout the paper that the user has access to a full estimated covariance matrix of cell estimates. However, such detailed information is often not available for secondary analyses, and in fact even cell deffs may not be available, as pointed out by Molina. In the latter case, Rao and Scott (1987) showed that an F statistic used in GLIM for testing a nested hypothesis, such as $\gamma = 0$ given the model (2.12), is asymptotically valid whenever the covariance matrix of cell estimates, \hat{V} , is proportional to the multinomial covariance matrix, \hat{P} . The F -test, however, is less powerful than the Rao-Scott tests, unless the denominator degrees of freedom are high. In the latter case, the F test might work well even if the condition $\hat{V} \propto \hat{P}$ is not satisfied (see Rao and Scott 1987, p. 392).

For the data in Table 1, $F = 6.63$ for testing $\gamma = 0$ given the model (2.12), which is not significant at the 5% level compared to $F_{1,3}(0.05) = 10.01$, the upper 1% of the F distribution with 1 and 3 degrees of freedom (d.f.). On the other hand, the Wald test W_1 and the Rao-Scott test, both requiring detailed information on the estimated covariance matrix, are significant at the 1% level compared to $\chi^2_1(0.01) = 6.63$. The F -test, therefore, appears to be less powerful here since the denominator d.f. is only 3. Molina's proposed test is, in fact, equal to F , but he was treating F as a χ^2 variable with 1 d.f. which may not be valid due to small denominator d.f.

The GLIM method does not provide a statistic for testing the goodness-of-fit of a model. Some information on the design effects is necessary for getting a valid test of goodness-of-fit.

(iii) Response to comments of C.J. Skinner

Skinner noted that the test of equality of two logistic regression models in Section 4 might be formulated as a test of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987), using dummy x -variables. The framework of Roberts, Rao and Kumar, however, assumes one fixed sample size n whereas in Section 4 we have two fixed sample sizes n_1 and n_2 for the two time periods. As a result, their results would need careful modification in order to be applicable to the present case of test of equality of two logistic regression models. Moreover, the dummy variable approach would involve the determination of estimates of $2s$ parameters iteratively, whereas the approach in Section 4 requires two iterative solutions, each involving only s parameters. Thus, the dummy variable approach could lead to convergence problems if s is not small.

We treated WLS with singular covariance matrices separately in Section 2 since the logit-type models in the remaining sections do not involve singular covariance matrices. WLS can also be applied to logit-type models but the resulting estimators and associated Wald tests may be unstable if the degrees of freedom associated with the estimated covariance matrix, \hat{V}_p , are low (criterion C3 of Skinner). The six criteria proposed by Skinner for comparing WLS and PML are very useful. We prefer PML mainly on the basis of criterion C3. Regarding the relative efficiency of WLS and PML estimators under complex designs, no general results are available, but WLS estimators are not likely to be significantly more efficient (and in fact, may be less efficient) if the degrees of freedom associated with the estimated covariance matrix are low. Clearly, further research on the relative efficiency of WLS and PML estimators would be useful.

ADDITIONAL REFERENCES

- RAO, J.N.K., and COLIN, D. (1988). Fitting dose-response models and hypothesis testing in teratological studies. Technical Report No. 116, Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa, Ottawa, Ontario.

Simultaneous Confidence Intervals for Proportions Under Cluster Sampling

D. ROLAND THOMAS¹

ABSTRACT

The paper describes a Monte Carlo study of simultaneous confidence interval procedures for $k > 2$ proportions, under a model of two-stage cluster sampling. The procedures investigated include: (i) standard multinomial intervals; (ii) Scheffé intervals based on sample estimates of the variances of cell proportions; (iii) Quesenberry-Hurst intervals adapted for clustered data using Rao and Scott's first and second order adjustments to X^2 ; (iv) simple Bonferroni intervals; (v) Bonferroni intervals based on transformations of the estimated proportions; (vi) Bonferroni intervals computed using the critical points of Student's t . In several realistic situations, actual coverage rates of the multinomial procedures were found to be seriously depressed compared to the nominal rate. The best performing intervals, from the point of view of coverage rates and coverage symmetry (an extension of an idea due to Jennings), were the t -based Bonferroni intervals derived using log and logit transformations. Of the Scheffé-like procedures, the best performance was provided by Quesenberry-Hurst intervals in combination with first-order Rao-Scott adjustments.

KEY WORDS: Simultaneous inference; Complex surveys; Monte Carlo.

1. INTRODUCTION

Survey results are often presented as estimated proportions (or percentages) of population units belonging to two or more distinct categories. Examples include many sociological studies (see for example Black and Myles 1986), marketing studies and opinion polls. As noted by Fitzpatrick and Scott (1987), inference on category proportions is often based on single binomial confidence intervals, even when more than two category proportions are being examined. This paper describes a study of several procedures for constructing simultaneous confidence intervals for the proportions π_i , $i = 1, \dots, k$, of population units belonging to each of k distinct categories, using data from a two-stage cluster sample. Standard simultaneous confidence interval procedures for categorical data problems, reviewed by Hochberg and Tamane (1987), are based on the assumption of multinomially distributed sample counts, and are thus appropriate for data from simple random samples. When the data have been collected using sample survey designs that involve clustering, standard procedures are likely to perform poorly, as is the case when standard multinomial based tests are applied to data from complex sample surveys. In the latter case, it has been shown by many workers that clustering can lead to unacceptably high Type I error rates (see, for example, Fellegi 1980; Rao and Scott 1979, 1981; Holt, Scott and Ewing 1980). For simultaneous confidence intervals, therefore, it is natural to expect that clustering will lead to coverage probabilities that are lower than multinomial theory indicates.

Estimation of simultaneous confidence intervals (SCI's) is an important adjunct to hypothesis testing. The present study thus represents a natural follow-up to Thomas and Rao's (1987) investigation of test statistics for the simple goodness of fit problem, under

¹ D. Roland Thomas, School of Business, Carleton University, Ottawa, Ontario, K1S 5B6.

simulated cluster sampling. In this paper, adaptations of the standard SCI procedures are proposed, and their performance in small samples is evaluated using Monte Carlo techniques.

The cluster sampling model that is used in the Monte Carlo study is described in Section 2, and the SCI procedures to be examined are presented in Section 3. In Section 4, the design of the Monte Carlo experiment is described, together with procedures for evaluating confidence interval performance. The main results of the study are presented in Sections 5 through 7, followed in Section 8 by some final conclusions and recommendations.

2. THE CLUSTER SAMPLING MODEL

This investigation will focus on two-stage sampling in which a k -category sample of m units is drawn independently from each of r sampled clusters.

For a sample of size $n = mr$, let $\mathbf{m} = (m_1, \dots, m_{k-1})'$ represent the category counts for the whole sample, where $m_k = n - \sum_{i=1}^{k-1} m_i$. In terms of proportions, let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_{k-1})' = \mathbf{m}/n$ be the vector of category proportions for the full sample. Further, define $\pi = E(\hat{\pi})$, where E denotes expectation under a suitable model of cluster sampling, and let V/n represent the $(k-1) \times (k-1)$ covariance matrix of $\hat{\pi}$. Following Rao and Scott (1981), the ordinary design effect for the linear combination $\mathbf{c}'\hat{\pi}$ of category proportions is $\mathbf{c}'V\mathbf{c}/\mathbf{c}'P\mathbf{c}$, where P is n times the covariance matrix of $\hat{\pi}$ under multinomial sampling, *i.e.*, $P = \text{diag}(\pi) - \pi\pi'$, and \mathbf{c} is a vector of dimension $k-1$. The largest design effect taken over all possible linear combinations is given by the largest eigenvalue of the design effect matrix $D = P^{-1}V$. The eigenvalues of D , denoted in decreasing order by $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$, were termed generalized design effects by Rao and Scott (1981), and provide a quantitative summary of the variance inflation associated with a particular design, relative to simple random sampling. Under the multinomial distribution, corresponding to simple random sampling from large populations, $\lambda_j = 1 \forall j$. Designs involving clustering usually yield generalized design effects greater than one on the average, *i.e.*, $\bar{\lambda} = \sum_{j=1}^{k-1} \lambda_j / (k-1) > 1$. Furthermore, studies of real survey data (Hidioglou and Rao 1987; Rao and Thomas 1988) reveal significant variation among the λ_j 's. This is conveniently represented by their coefficient of variation, given by

$$a = \left(\sum_{j=1}^{k-1} \lambda_j^2 / [(k-1)\bar{\lambda}^2] - 1 \right)^{1/2}. \quad (1)$$

A suitable model of cluster sampling must therefore be capable of generating generalized design effects such that $\bar{\lambda} > 1$ and $a > 0$.

Brier (1981) proposed a model of two-stage cluster sampling in which individual clusters are represented by vectors of category probabilities, $\mathbf{p}_\ell = (p_{\ell 1}, p_{\ell 2}, \dots, p_{\ell, k-1})'$, $\ell = 1, \dots, r$, where for each cluster, $p_{\ell k} = 1 - \sum_{i=1}^{k-1} p_{\ell i}$. Each \mathbf{p}_ℓ was independently drawn from a Dirichlet distribution with mean π , *i.e.* $E(\mathbf{p}_\ell) = \pi$, and second stage sampling of the m units per cluster was multinomial, conditional on the realized value of \mathbf{p}_ℓ for that cluster. Let the vector of counts for each cluster be $\mathbf{m}_\ell = (m_{\ell 1}, \dots, m_{\ell, k-1})'$, with $m_{\ell k} = m - \sum_{i=1}^{k-1} m_{\ell i}$. Thus for the full sample, $\mathbf{m} = \sum_{\ell=1}^r \mathbf{m}_\ell$, and in terms of proportions, $\hat{\pi} = \sum_{\ell=1}^r \hat{\pi}_\ell$, where $\hat{\pi}_\ell = \mathbf{m}_\ell/m$. Brier (1981) showed that under this model, $E(\hat{\pi}) = \pi$ and $V(\hat{\pi}) = dP/n$, *i.e.*, the covariance matrix of $\hat{\pi}$ is proportional to the multinomial covariance matrix, with the constant of proportionality $d > 1$. Under this model, the design effect matrix is given by $D = dI_{k-1}$, where I_{k-1} is the identity matrix of order $k-1$. Thus $\lambda_i = d \forall i$, so that $\bar{\lambda} = d$ and $a = 0$. Brier's model can therefore represent variance inflation ($\bar{\lambda} > 1$), but cannot

represent the unequal generalized design effects encountered in practice. Thomas and Rao (1987) used an extension of Brier's model in which the first stage p_i 's are sampled independently from a mixture of two Dirichlet distributions, representing a population composed of two distinct classes of clusters. This model, which is a special case of that proposed by Rao and Scott (1979), generates one distinct and $k - 2$ equal eigenvalues, with $\bar{\lambda}$ and a being explicit functions of the Dirichlet parameters. This greatly facilitates the design of the Monte Carlo study by allowing for convenient control of the values of the clustering measures $\bar{\lambda}$ and a . Since it satisfies the basic requirements outlined above ($\bar{\lambda} > 1, a > 0$), Thomas and Rao's (1987) model will be used in this study.

3. SIMULTANEOUS CONFIDENCE INTERVAL PROCEDURES

3.1 Scheffé Intervals

A standard Scheffé argument, based on the asymptotically exact probability statement

$$P\left(n(\hat{\pi} - \pi)' \hat{V}^{-1} (\hat{\pi} - \pi) \leq \chi_{k-1}^2(\alpha)\right) = 1 - \alpha \quad (2)$$

leads to simultaneous confidence intervals for linear combinations, $\ell' \pi$, of the category probabilities, where ℓ is a vector of dimension $(k - 1)$. Appropriate choices of ℓ then yield SCI's on the individual cell probabilities given by

$$\pi_i \in \left\{ \hat{\pi}_i \pm (\hat{v}_{ii})^{1/2} (A/n)^{1/2} \right\}, i = 1, \dots, k, \quad (3)$$

where $A = \chi_{k-1}^2(\alpha)$ is the upper α percent point of a chi-squared distribution on $k - 1$ degrees of freedom, and \hat{v}_{ii} is the i^{th} diagonal element of a consistent estimator of V (as $r \rightarrow \infty$) given by

$$\hat{V} = \frac{n}{r(r-1)} \sum_{\ell=1}^r (\hat{\pi}_{\ell} - \hat{\pi}) (\hat{\pi}_{\ell} - \hat{\pi})'. \quad (4)$$

Note that when the endpoint of an interval lies outside $[0, 1]$, definition (3) must be modified by truncating the endpoint to 0 or 1 as appropriate. For multinomial sampling, \hat{v}_{ii} can be replaced by $\hat{\pi}_i (1 - \hat{\pi}_i)$, in which case the Scheffé intervals reduce to those proposed by Gold (1963). The latter will be referred to as Scheffé-Gold intervals. The Scheffé intervals of equation (3) will be conservative, *i.e.*, will have coverage exceeding $(1 - \alpha)$ asymptotically since they make use of only a finite number of the available ℓ directions (see Miller 1981, page 63). In fact, they will become very conservative as k increases, as can be shown using the following argument due to Goodman (1965). The coverage of the Scheffé intervals is equal to one minus the probability of occurrence of at least one of the events $\{(\hat{\pi}_i - \pi_i)^2 / (\hat{v}_{ii}/n) > \chi_{(k-1)}^2(\alpha)\}$, $i = 1, \dots, k$; since the random variables $(\hat{\pi}_i - \pi_i)^2 / (\hat{v}_{ii}/n)$ each have chi-squared distributions on one degree of freedom asymptotically, the probability of each individual event can be evaluated. Using the Bonferroni inequality, lower bounds for the coverage can then be obtained; for a nominal coverage of 95% with $k = 3, 5, 8$ and 12 , these bounds are .9571, .9896, .9986 and .9999 respectively.

3.2 Modified Quesenberry-Hurst Intervals

Under the assumption of multinomial sampling, Quesenberry and Hurst (1964) solved the large sample probability statement

$$P\left\{X^2 = n \sum_{i=1}^k \frac{(\hat{\pi}_i - \pi_i)^2}{\pi_i} \leq A\right\} = 1 - \alpha \quad (5)$$

for the cell probabilities π_i , to get the SCI's

$$\pi_i \in \left\{ \frac{\hat{\pi} + A/2n \pm (A/n)^{1/2} [\hat{\pi}_i (1 - \hat{\pi}_i) + A/4n]^{1/2}}{1 + A/2n} \right\}. \quad (6)$$

Under multinomial sampling, these intervals are asymptotically equivalent to Scheffé and Scheffé-Gold intervals, and will therefore exhibit similar asymptotic conservativeness.

Quesenberry-Hurst (Q-H) intervals can be adapted for use with clustered survey data using the first and second order corrections to the distribution of X^2 proposed by Rao and Scott (1981). Corresponding first and second order SCI's are obtained by replacing A in equation (3) by

$$A^{(1)} = \hat{\lambda}A \text{ and } A^{(2)} = \hat{\lambda}(1 + \hat{a}^2) \chi_v^2(\alpha) \quad (7)$$

respectively, where $v = (k - 1)/(1 + \hat{a}^2)$ and $\hat{\lambda}$, an estimate of the mean of the generalized design effects, is given by (Rao and Scott, 1981)

$$\hat{\lambda} = (k - 1)^{-1} \sum_{i=1}^k (1 - \hat{\pi}_i) \hat{d}_i, \quad (8)$$

where \hat{d}_i , $i = \dots, k$ is an estimated cell design effect given by $\hat{d}_i = \hat{v}_{ii}/\hat{\pi}_i (1 - \hat{\pi}_i)$. The coefficient of variation, a , is estimated by replacing $\bar{\lambda}$ in equation (1) by $\hat{\lambda}$, and $\sum \lambda_i^2$ by the estimate $\sum \hat{\lambda}_i^2 = \sum \sum \hat{v}_{ij}^2 / \hat{\pi}_i \hat{\pi}_j$. It turns out (see Thomas 1989) that the second order modified intervals are unnecessarily conservative, so that only the first-order modified Q-H intervals will be discussed in the remainder of the paper.

3.3 Simple Bonferroni Intervals

Since (loosely speaking) each $\hat{\pi}_i$ is asymptotically $N(\pi_i, v_{ii}/n)$, the intervals

$$\pi_i \in \left\{ \hat{\pi}_i \pm (\hat{v}_{ii}/n)^{1/2} z_{\alpha'/2} \right\}, \quad (9)$$

will have large sample coverage at least $(1 - \alpha)$ by the Bonferroni inequality, where $\alpha' = \alpha/k$ and $z_{\alpha'/2}$ is the upper $\alpha'/2$ percent point of the standard normal distribution. Intervals (9) are equivalent to Scheffé intervals with A in equation (3) replaced by $A^{(3)} = \chi_1^2(\alpha')$. As noted

by Goodman (1965), they will be shorter than Scheffé intervals for the usual values of α and k ; e.g., $\alpha = 1\%$, 5% , or 10% . Goodman's (1965) multinomial Bonferroni intervals are given by equation (9) with \hat{v}_{ii} replaced by $\hat{\pi}_i (1 - \hat{\pi}_i)$. All endpoints of simple Bonferroni intervals that lie outside $[0, 1]$ will be truncated to 0 or 1 as appropriate.

3.4 Transformed Bonferroni Intervals

For suitably smooth g , $g(\hat{\pi}_i)$ will be asymptotically $N(g(\pi_i), [g'_i(\pi_i)]^2 v_{ii}/n)$, where $g'_i(\pi_i)$ denotes the partial derivative $\partial g(\pi_i)/\partial \pi_i$ evaluated at π_i . Bonferroni intervals can then be obtained by inverting corresponding intervals on the $g(\pi_i)$'s, giving

$$\pi_i \in \left\{ g^{-1}(g(\hat{\pi}_i) \pm g'_i(\hat{\pi}_i) (\hat{v}_{ii}/n)^{1/2} z_{\alpha'/2}) \right\}. \quad (10)$$

Three g functions will be investigated: the square root $g_1(\pi_i) = \pi_i^{1/2}$ (previously investigated by Bailey 1980, for the case of multinomial sampling); the natural logarithm $g_2(\pi_i) = \ln(\pi_i)$; and the logit $g_3(\pi_i) = \ln(\pi_i/(1 - \pi_i))$. Interval endpoints that lie outside $[0, 1]$ will again be truncated to 0 or 1 as necessary.

Transformed Bonferroni intervals based on a jackknifed estimator of the variance of $g(\hat{\pi})$ have also been examined (see Thomas 1989). It was found that there is little advantage to using jackknifed variance estimates; Taylor series variance estimates are therefore recommended for their simplicity. Intervals based on jackknife variance estimates will not be considered further in this paper.

3.5 Variants of the Above Intervals

Scheffé Intervals: Following Thomas and Rao (1987), Scheffé intervals can be modified by replacing the critical constant A in equation (3) by $A^{(4)} = (k - 1)(r - 1)(r - k + 1)^{-1} F_{(k-1), (r-k+1)}(\alpha)$, where $F_{(k-1), (r-k+1)}(\alpha)$ is the upper α percent point of an F distribution on $(k - 1)$ and $(r - k + 1)$ degrees of freedom.

Quesenberry-Hurst Intervals: Variants of the modified Quesenberry-Hurst (Q-H) intervals can also be defined, corresponding to the F forms of the first and second order corrected test statistic proposed by Thomas and Rao (1987). These again turn out to be conservative, and will not be considered further.

Bonferroni Intervals: Heuristic arguments (see the appendix to Thomas and Rao 1987) suggest that the simple Bonferroni intervals can be improved by replacing $z_{\alpha'/2}$ in (9) by $t_{r-1}(\alpha'/2)$, the upper $\alpha'/2$ percentage point of Student's t distribution on $r - 1$ degrees of freedom. This strategy will also be applied to the transformed Bonferroni intervals.

4. THE DESIGN OF THE MONTE CARLO STUDY

4.1 Parameters and Random Numbers

The parameters to be controlled are: (i) the nominal coverage level $(1 - \alpha)$ of the SCI; (ii) π , the model probability vector; (iii) k , the number of categories; (iv) r , the number of sample clusters; (v) m , the number of units drawn from each sampled cluster; (vi) $\bar{\lambda}$, the mean of the generalized design effects (eigenvalues); (vii) a , the coefficient of variation of the generalized design effects. The nature and degree of clustering is represented by the pair $(\bar{\lambda}, a)$ as follows: (a) multinomial sampling ($\bar{\lambda} = 1, a = 0$); (b) constant design effect clustering ($\bar{\lambda} > 1, a = 0$); (c) non-constant design effect clustering ($\bar{\lambda} > 1, a > 0$).

Individual Monte Carlo experiments were run for particular combinations of k , $\bar{\lambda}$, a and r_{max} , the latter being the maximum number of clusters generated in one computer run. Most experiments were run at two values of $\bar{\lambda}$, namely 1.5 and 2.0, two values of a , namely $a = 0$ (constant design effects) and $a > 0$ (one level of non-constant design effects), for equiprobable categories ($\pi_i = 1/k, i = 1, \dots, k$). Three values of k ($k = 3, 5, 8$) were initially selected to cover the range of numbers of categories commonly encountered in goodness-of-fit tests. An additional run was subsequently done for the case $k = 12$, $\bar{\lambda} = 2$ and $a > 0$ to check on the range of applicability of the results. The number of units per cluster was set at $m = 10$ for $k = 3, 5$ and 8 , and at $m = 20$ for $k = 12$. Preliminary investigations showed coverage rates to be insensitive to the value of this parameter. For comparability of results over k , the non-zero settings of a were selected to make a/a_{max} the same for each selected value of k , where $a_{max} = (k - 2)^{1/2}$ is the maximum possible value of a . For $k = 5$, the non-zero value of a was set at 0.5, which is typical of the values encountered in practice, *e.g.*, $\hat{a} = 0.43$ for $k = 5$, as reported by Rao and Thomas (1988).

The initial focus on equiprobable categories allowed for a cost effective assessment of the influence of k , $\bar{\lambda}$ and a on coverage rates, and eliminated many of the possible SCI variants from further consideration. Additional experiments reported in Section 7 show that the procedures that passed this initial screening can in fact be applied when the cell probabilities are markedly unequal. Vectors of unequal probabilities were confined to the class $\pi(k, q, \phi)$, defined by the elements $\pi_i = \phi, i = 1, \dots, q$ and $\pi_i = (1 - q\phi)/(k - q), i = q + 1, \dots, k$.

For details of the generation of the random clusters from the mixture Dirichlet multinomial distribution, the reader is referred to Thomas and Rao (1987). Each Monte Carlo experiment consisted of 1000 sets of up to 100 independent clusters, grouped into nested subsets. All SCI procedures were applied in turn to each subset, using two nominal coverage levels (95% and 90%), thus improving the precision of comparisons between procedures at the same parameter settings, and between the same SCI procedures for different numbers of clusters. Most of the results presented will be for 95% nominal coverage; trends for 90% coverage were found to be qualitatively similar.

4.2 Evaluation Procedures

The percentage of Monte Carlo trials for which at least one of the k confidence intervals fails to cover the true parameter value is reported, and used for a preliminary screening of the main SCI procedures. This is a measure of the family error rate, which is equivalent to the actual significance level of the SCI when the latter is viewed as a test of goodness-of-fit. The family error rate, which will be referred to in this paper as the total error rate ER_T , is used in place of the more commonly reported actual coverage rate (equal to one hundred percent minus the total error rate) because it can be conveniently split into two one-sided rates which will provide information on the symmetry or 'unbiasedness' of each SCI procedure. Jennings (1987) argued that coverage rates alone can provide a misleading assessment of single parameter confidence interval procedures, and recommended that the number of times that an interval falls above and below the true parameter value should be separately reported. In this paper, Jennings' suggestion has been adapted to simultaneous confidence intervals on $\pi_i, i \in I$, where I is the index set $\{1, \dots, k\}$, by counting the number of Monte Carlo trials for which:

- (a) more intervals fall above their corresponding $\pi_i, i \in I$, than fall below;
- (b) more intervals fall below their corresponding $\pi_i, i \in I$, than fall above;
- (c) the same number (> 0) of intervals fall above their corresponding $\pi_i, i \in I$, as fall below.

Upper and lower error rates are then defined as $ER_U = [n_a + (n_c/2)]/N_t$ and $ER_L = [n_b + (n_c/2)]/N_t$, respectively, where N_t represents the number of Monte Carlo trials, and n_a , n_b and n_c denote the counts (a) through (c), respectively. The sum of ER_U and ER_L is clearly equal to the total error rate, ER_T . These one-sided error rates will be used to compare SCI procedures whose overall error rates are acceptably close to the nominal rate α , over a range of parameter settings and cluster strengths. Average interval lengths and corresponding standard errors have also been computed, and will be used as final discriminators in the selection of the recommended procedures.

5. A SUMMARY OF RESULTS FOR TOTAL ERROR RATES

All results in this section are given in terms of the total error rate ER_T , defined in Section 4. For lack of space, tables are presented only for the case of unequal design effects, ($a > 0$), with $\bar{\lambda} = 2$. More detailed results are given in Thomas (1989). In interpreting the tabulated results, it should be noted that for 1000 Monte Carlo trials, binomial standard errors of point estimates of true ER_T 's having magnitudes 5%, 10% and 20% are 0.7%, 0.9% and 1.3% respectively. As a general rule deviations from nominal rates, and differences between the error rates of different SCI procedures will be noted only when they are large enough to have practical significance, and exceed their Monte Carlo standard errors by a factor of at least two.

5.1 Multinomial Procedures

Results for multinomial intervals will only be summarized here; for details see Thomas (1989). Under cluster sampling, error rates for Goodman's Bonferroni intervals (see equation (9) with \hat{p}_{ii} replaced by $\hat{\pi}_i(1 - \hat{\pi})$) are unacceptably high except for values of $\bar{\lambda}$ close to 1, *i.e.*, unless the effect of clustering is small. The Scheffé-Gold and multinomial Quesenberry-Hurst intervals, on the other hand, can yield error rates that are close to the nominal value in certain cases, whenever their inherent conservativeness balances the error inflating effects of clustering (see also Andrews and Birdsall 1988). Unfortunately, this is not always the case; both procedures can display inflated error rates ($ER_T \geq 2\alpha$) for realistic combinations of category numbers and clustering strengths.

Multinomial procedures should therefore not be used with complex survey data. Procedures are clearly required that directly account for the clustering, and provide good coverage for the required number of categories, over a wide range of clustering conditions.

5.2 The Scheffé Procedures

Total error rates for the χ^2 -based Scheffé procedure of equation (3) and its F -based variant are summarized in Table 1 as functions of r , for the case $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$. More detailed graphs are given in Thomas (1989).

For the values of k studied, ER_T for the χ^2 -based Scheffé procedure of equation (3) increases rapidly as the number of clusters decreases, so that it should never be used for small numbers of clusters. The F -based variant, on the other hand, keeps ER_T reasonably close to or below $\alpha = 5\%$ for all r . As r increases, ER_T for F -based Scheffé remains fairly constant for the case $k = 3$, but becomes increasingly conservative for $k \geq 5$, as does the χ^2 version. These empirical trends with varying r can be explained in terms of two competing effects. As r increases, error rates for both procedures approach their asymptotic levels which are bounded above by 4.29%, 1.04% and 0.14%, for $k = 3, 5$ and 8 respectively (see Section 3.1).

Table 1
Total Error Rates for Scheffé and Modified Q-H Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

<i>k</i>	<i>a</i>	<i>r</i>	Total Error Rate (<i>ER_T</i>)		
			Scheffé (χ^2 based)	Scheffé (<i>F</i> based)	Modified Q-H (first order)
3	.29	15	9.2	5.9	5.0
3	.29	30	5.7	4.7	5.1
3	.29	50	5.4	5.0	5.4
5	.5	15	8.8	5.2	4.3
5	.5	30	4.0	3.0	2.7
5	.5	50	2.5	2.0	2.0
8	.71	15	12.7	7.4	2.4
8	.71	30	4.2	3.0	2.7
8	.71	50	2.7	1.6	2.5
8	.71	100	0.8	0.7	2.3

As *r* decreases, however, the conservativeness of the Scheffé procedures (for $k \geq 5$) will be increasingly swamped by the effects of increasing non-normality of the estimated proportions, $\hat{\pi}$. For the *F*-based version, the inflation in error rate due to non-normality is less than for the chi-squared version of equation (3), with the result that *ER_T* for the *F*-based version never seriously exceeds the nominal 5% rate. For moderate levels of clustering ($\bar{\lambda} = 1.5$), the behaviour of the *F*-based procedure is qualitatively similar to that described above for the case $\bar{\lambda} = 2$. From the point of view of total error rate, therefore, the *F*-based Scheffé procedure is useable over a wide range of clustering situations, though its possible conservativeness is a disadvantage.

5.3 Modified Quesenberry-Hurst Intervals

Total error rates for the first order modified Quesenberry-Hurst (Q-H) procedure of Section 3.2 are also shown in Table 1 for $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$.

Total error rates are close to or below the nominal 5% for all combinations of *r* and *k* shown. For moderate to large numbers of clusters ($r \geq 30$), error rates for $k = 5$, and 8 are very similar, being approximately one half of the nominal rate (true also when $k = 12$). For the case of constant design effects (see Thomas 1989), error rates for first order modified Q-H intervals are conservative for $k \geq 5$, particularly for large *r*. The absence of this Scheffé-like conservativeness for the more realistic case of unequal design effects shown in Table 1 can again be explained using the argument of Section 3.1. From equation (6), it is easily seen that the asymptotic coverage of the first-order modified Q-H intervals is given by one minus the probability that at least one of the random variables $(\hat{\pi}_i - \pi_i)^2 / (\bar{\lambda} \pi_i (1 - \pi_i) / n)$, $i = 1, \dots, k$, will exceed the critical value $\chi^2_{k-1}(\alpha)$ asymptotically. When $a > 0$, these individual random variables will not all be asymptotically distributed as chi-squared on one degree of freedom, so that the bound of Section 3.1 does not apply. The true bound on the error rate will be inflated since at least one of the random variables $(\hat{\pi}_i - \pi_i)^2 / (\bar{\lambda} \pi_i (1 - \pi_i) / n)$ will be stochastically larger than $(\hat{\pi}_i - \pi_i)^2 / (\nu_H / n)$, whenever $a > 0$.

Trends for the case $\bar{\lambda} = 1.5$ are similar (Thomas 1989). Overall, the results show that from the point of view of total error rates, first-order modified Q-H intervals provide a safe but somewhat conservative SCI procedure under realistic clustering conditions.

5.4 Simple Bonferroni Intervals

Total error rates for the simple Bonferroni intervals given by equation (9) are summarized in Table 2 for the case $\alpha = 5\%$, $\bar{\lambda} = 2$, $a > 0$, and $k = 3, 5$ and 8 . Also shown are corresponding error rates for the t -based variants described in Section 3.5.

From Table 2, it is evident that the error performance of both sets of SCI's is poor, both showing a strong tendency to high error rates for small to medium numbers of clusters when k , the number of categories, is five or more. Using critical values of Student's t distribution to compensate for the variability in the estimated variances of the category proportions clearly has the effect of generally lowering error rates. As can be seen from Table 2, however, this strategy is unable to prevent significant error rate inflation in the t -based intervals as the number of clusters decreases, except when $k = 3$. The trend to inflated error rates for small numbers of clusters (for both z and t -based intervals), is due to the increasing non-normality of the $\hat{\pi}_i$'s with decreasing r . This trend gets progressively more severe as k increases, which is to be expected since non-normality will become more pronounced, for a given value of r , as the values of the π_i 's get smaller. This is precisely what happens with increasing k in the case under study, for which $\hat{\pi}_i = 1/k \forall i$.

When $k = 3$, error rates for the t -based procedure are essentially constant, and close to the nominal level. For $k = 8$, on the other hand, ER_T varies from close to 20% at $r = 15$ to approximately 8% at $r = 100$. From Table 2, and other results not shown, it appears that for $k \geq 8$, simple t -based intervals approach their Bonferroni limits very slowly as $r \rightarrow \infty$. Also, for $k \leq 5$, error rates are close to the nominal level for moderate to large numbers of clusters ($r \geq 40$). Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are consistent with the above. From the point of view of total error rates (or equivalently of coverage rates), it is clear that simple t -based Bonferroni intervals are useable in practice over a range of realistic clustering situations only if $k \leq 5$ and $r \geq 40$.

Table 2
Total Error Rates for z and t -Based Simple Bonferroni Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

k	a	r	Total Error Rate (ER_T)	
			z -based	t -based
3	.29	15	10.0	5.6
3	.29	30	6.3	4.9
3	.29	50	6.5	5.5
5	.50	15	15.0	9.7
5	.50	30	8.8	7.2
5	.50	50	7.2	5.5
8	.71	15	29.6	19.1
8	.71	30	15.0	11.0
8	.71	50	11.5	9.8
8	.71	100	8.1	7.8

5.5 Transformed Bonferroni Intervals

The more detailed results given in Thomas (1989) demonstrate that the problem of error rate inflation exhibited by simple z -based Bonferroni intervals is not solved by the use of transformations alone. All three transformed z -based intervals again display severely inflated error rates for small to medium numbers of clusters. Fortunately, the effect of transformations on the t -based Bonferroni intervals is very different, as can be seen from the results summarized in Table 3.

For $k = 3, 5$ and 8 , error rates for the log and logit intervals are close to the nominal 5% for all r values shown, with the logit intervals yielding slightly lower rates than the log intervals (see the footnote to Table 3). The t -based square root intervals, on the other hand, exhibit the undesirable characteristic of error rate inflation for small r , when $k \geq 8$; they will not be considered further. For large numbers of categories ($k = 12$), both log and logit intervals do exhibit some error rate inflation for intermediate numbers of clusters ($r = 30$). This is not a serious drawback, however, as this number of categories is rarely encountered in practice. Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are generally similar to those described above.

It thus appears that for the ranges of $k, r, \bar{\lambda}$ and α that are likely to be encountered in practice, log and logit transformations (which reduce the non-normality in $\hat{\pi}$) used in combination with t -based critical values (which compensate for the variability in the estimated variances) do yield intervals that provide the desired degree of control. These intervals will be explored further in Section 6 in terms of the symmetry of their error rates.

Table 3
Total Error Rates¹ for t -based Transformed Bonferroni Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

			Total Error Rate (ER_T)		
k	α	r	t -based Transformed Bonferroni		
			Square Root	Log	Logit
3	.29	15	4.5	4.6	3.3
3	.29	30	3.6	4.0	3.5
3	.29	50	4.6	5.6	4.1
5	.5	15	6.4	4.7	4.6
5	.5	30	4.6	4.2	3.5
5	.5	50	4.3	4.5	4.0
8	.71	15	12.0	5.9	5.2
8	.71	30	6.2	6.6	5.2
8	.71	50	5.9	5.4	5.2
8	.71	100	4.9	3.9	4.2
12	.91	15	17.0	6.7	6.5
12	.91	30	12.9	10.1	10.2
12	.91	50	8.2	6.5	6.3

¹ For $k = 8$ and $r = 50$, the correlation between ER_T estimates for log and logit intervals is 0.92. Assuming this is typical for all r and k , the Monte Carlo standard error of the difference in log and logit error rates is approximately 0.3%.

Table 4

Percentage Asymmetry (PER_U)¹ in the Total Error Rate for the Viable Procedures;
 $a > 0^2$, $r = 50$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

α	k	$\bar{\lambda}$	$PER_U = (ER_U/ER_T) \times 100\%$			
			Scheffé (F -based)	Modified Q-H (first order)	t -based Bonferroni (log)	Bonferroni (logit)
5%	5	1.5	19.2	58.7	61.0	48.9
5%	5	2.0	0.0	45.0	61.1	48.8
5%	8	1.5	0.0	63.2	67.5	56.8
5%	8	2.0	0.0	65.2	64.9	49.0
5%	12	2.0	0.0	46.9	53.8	51.6
10%	5	1.5	16.3	49.4	59.2	48.4
10%	5	2.0	6.1	50.0	61.8	48.6
10%	8	1.5	0.0	60.7	67.3	55.8
10%	8	2.0	0.0	65.6	60.7	50.0
10%	12	2.0	0.0	47.5	56.0	51.4

¹ For $k = 8$, $\bar{\lambda} = 2$ and $\alpha = 5\%$, the correlation between PER_U estimates for log and logit intervals is 0.82. Assuming this is typical, Monte Carlo standard errors for differences in log and logit PER_U 's are approximately 4% and 3% for $\alpha = 5\%$ and 10%, respectively.
² For values of a for specific k , see Table 3.

6. ERROR RATE SYMMETRIES FOR THE VIABLE PROCEDURES

This section presents results on error rate symmetry based on the decomposition of the total error rate ER_T into its two additive components ER_U and ER_L , as described in Section 4. The measure used in the tables is $(ER_U/ER_T) \times 100\%$, i.e., the upper error rate expressed as a percentage of the total error rate. It will be denoted PER_U . A symmetric SCI will have an empirical PER_U that is close to 50%; a PER_U that is greater (less) than 50% will indicate an increased probability of non-coverage due to intervals lying above (below) their respective π_i 's. For values of percentage symmetry between 50% and 80%, 95% confidence intervals on the true PER_U are approximately $(PER_U \pm 14)\%$ and $(PER_U \pm 10)\%$ for total error rates of 5% and 10% respectively.

6.1 Modified Scheffé and Quesenberry-Hurst Intervals

Percentage symmetry results for the F -based Scheffé and the first order Quesenberry-Hurst (Q-H) intervals are given in Table 4 for a selection of parameter values. It can be seen that the Scheffé procedure displays extreme asymmetry, making it an unattractive SCI. The first order modified Q-H procedure displays only moderate asymmetry, and is therefore the better of the two in practice.

The source of the asymmetry in the Scheffé intervals is again the non-normality of the untransformed $\hat{\pi}_i$'s. In particular, the fact that "small" $\hat{\pi}_i$'s generate "small" estimates of the variances ν_{ii} and hence shorter intervals (cf. the multinomial case where $\hat{\nu}_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)/n$, $i = 1, \dots, k$) increases the probability that non-covering intervals will lie below their respective π_i 's. This tendency to asymmetry will increase as the total error rate decreases, making the F -based Scheffé procedure particularly vulnerable to this effect. Since Scheffé intervals differ from simple Bonferroni intervals only through the critical constant used, asymmetry is also to be expected in the latter though it should not be as severe given that error rates for simple Bonferroni intervals are liberal. This is confirmed by study results, e.g., $PER_u = 4.9\%$ for simple t -based Bonferroni intervals when $r = 50$, $k = 8$ and $a = 0.71$.

6.2 *t*-Based Transformed Bonferroni Intervals

Table 4 also gives percentage symmetry results for *t*-based Bonferroni intervals based on the log and logit transformations. The results of the table suggest that logit intervals do provide more symmetric coverage than the log intervals, when *k* is in the range 5 to 8. Thus logit intervals might be considered preferable in practice to log intervals from the point of view of error rate symmetry.

7. UNEQUAL CELL PROBABILITIES

Table 5 presents results on total error rates and error rate symmetry under unequal cell probabilities for the *t*-based log and logit transformed Bonferroni procedures, together with results for the first order modified Q-H procedure. Results are tabulated for six sets of unequal probabilities, three for the case $k = 5, \bar{\lambda} = 2, a = 0.5$, namely $\pi(5, 3, .3), \pi(5, 2, .425)$ and $\pi(5, 1, .8)$, (see Section 4.1), and three for the case $k = 8, \bar{\lambda} = 2, a = 0.71$, namely $\pi(8, 3, .25), \pi(8, 2, .35)$ and $\pi(8, 1, .65)$. For each π vector, the remaining $k - q$ elements all equal 0.05. Results for equiprobable cells are also displayed in Table 5 for comparison.

It can be seen that deviations from equiprobability do affect total error rates for the modified Q-H procedure, particularly when $k = 8$. With the first element $\pi_1 = 0.65$ the total error rate of modified Q-H is close to its error rate under equiprobability. For the other two cases studied ($\pi_1 = \pi_2 = .35$, and $\pi_1 = \pi_2 = \pi_3 = 0.25$), total error rates are considerably lower, closer in fact to the modified Q-H results obtained for the constant design effect case (see Thomas 1989). This difference in total error rates occurs because the pattern of cell design effects is different for each set of unequal probabilities, though the pattern of generalized design effects (the λ 's) remains the same ($\lambda_1 = 2 + 2\sqrt{3}, \lambda_j = 2 - \sqrt{3}/3, j = 2, \dots, 7$ for $\bar{\lambda} = 2, a = \sqrt{2}/2 = .707$). When $\pi_1 = 0.65$, the cell design effects are $d_1 = 5.7, d_i = 1.82, i = 2, \dots, 8$.

Table 5
The Effect of Unequal Cell Probabilities on the Total Error Rates (ER_T) and Percentage Asymmetries (PER_U) of the Modified Q-H and Transformed Bonferroni Procedures;
 $r = 50, \bar{\lambda} = 2, a = 5\%, m = 10$

k $\pi(k,q,\phi)$		Procedures					
		Modified Q-H (first order)		t -based Bonferroni			
				(log)		(logit)	
		ER_T	PER_U	ER_T	PER_U	ER_T	PER_U
5	$\pi(5,1,0.8)$	3.2	7.3	5.6	75.9	4.4	62.5
5	$\pi(5,2,0.425)$	1.4	82.1	4.8	57.2	4.6	47.8
5	$\pi(5,3,0.3)$	1.5	76.7	4.2	51.2	3.9	38.5
5	equi-prob.	2.0	45.0	4.5	61.1	4.0	48.8
8	$\pi(8,1,0.65)$	2.7	63.0	6.3	68.3	5.4	55.6
8	$\pi(8,2,0.35)$	0.6	83.3	4.9	58.2	4.4	51.2
8	$\pi(8,3,0.25)$	0.7	100	5.2	68.2	4.6	63.1
8	equi-prob.	2.5	66.5	6.0	64.0	5.2	49.0

Use of a uniform adjustment factor ($\hat{\lambda}$) will thus seriously underestimate the variance of the first estimated cell probability, leading to inflation of the error rate of the modified Q-H procedure. That the nominal error rate $\alpha = 5\%$ is not exceeded is due to the inherent conservativeness of modified Q-H intervals in the constant design effect case (see Section 5.3). When $\pi_1 = \pi_2 = 0.35$, corresponding design effects are $d_1 = d_2 = 2.36$, $d_i = 1.97$, $i = 3, \dots, 8$. These are much closer to constant design effects ($d_i = 2.0$, $i = 1, \dots, 8$) hence the conservative behaviour of the intervals in this case. It can also be seen from Table 5 that conservative ER_T 's are associated with highly asymmetric error rates.

Despite the variation in cell design effects implied by the different probability vectors of Table 5, it can be seen that the transformed Bonferroni procedures exhibit very stable performance. Total error rates (for 50 clusters) are close to the nominal rate ($\alpha = 5\%$) for both log and logit intervals, and neither exhibits serious asymmetry. Total error rates corresponding to unequal probabilities do decrease with decreasing r over the range $r = 50$ to $r = 15$ when $k = 8$ (results not shown). Variations in ER_T are not severe, however; when $r = 15$ clusters the minimum rate for the cases examined is approximately 2%.

8. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In the search for procedures that take direct account of the survey design and that provide adequate control of error rates and error rate symmetry over a wide range of problem and clustering situations, Scheffé intervals based on estimated cell variances must be rejected: the chi-squared version of equation (3) on the grounds of poor error control, and the F -based version on the grounds of extreme asymmetry. Modifications to Quesenberry-Hurst intervals are somewhat conservative, though the version based on the first order Rao-Scott correction does provide a viable procedure. For Bonferroni intervals, the benefits of using critical points of the t -distribution instead of the standard normal are substantial. Even so, intervals based on $\hat{\pi}$ and its square root provide inadequate control of total error rates, particularly for small numbers of clusters when the distribution of $\hat{\pi}$ becomes increasingly non-normal. On the other hand, t -based Bonferroni intervals using both the log and logit transformations provide good control of total error rates and error rate symmetry, and are clearly superior to all other competing intervals. Both log and logit transformed intervals (t -based) also appear to provide good control of error rates and error rate symmetry when the cell probabilities are unequal, differing in the cases studied by a ratio (maximum to minimum) of up to sixteen. From the point of view of total error rates there is little to choose between the log and logit intervals, though error rates for the latter are consistently a little lower. Logit intervals are superior from the point of view of symmetry, however. Estimates of confidence interval lengths (detailed results not shown) also favour the logit intervals, despite their slightly lower error rates. For example, for the equiprobable case with $\alpha = 5\%$, $k = 5$, $\bar{\lambda} = 2$, $a = 0.5$ and $r = 50$, the average length of the confidence interval on π_1 (expressed as a 95% confidence interval) was $.1915 \pm .0014$ for the log-based interval, and $.1850 \pm .0014$ for the logit-based interval. For the case of unequal probabilities, with $\alpha = 5\%$, $k = 8$, $\bar{\lambda} = 2$, $a = 0.71$, $r = 50$, $\pi_1 = 0.65$ and $\pi_2 = 0.05$ (see Table 5), 95% confidence intervals for the average lengths of log and logit intervals were: for π_1 , $.2865 \pm .0012$ and $.2776 \pm .0011$, respectively; for π_2 , $.0806 \pm .0010$ and $.0789 \pm .0011$, respectively.

Before final recommendations are made, it is necessary to consider possible limitations imposed by the design of the Monte Carlo study. A potentially limiting feature is the use of a single specific sampling design, namely two-stage cluster sampling with SRS at the second

stage, given that practitioners will encounter data collected using a range of survey designs that might include stratification and multiple levels of unit selection. For large samples, the relevant distribution theory requires knowledge only of first and second moments, assuming that a suitable central limit theorem applies (see for example Rao and Scott 1981). This study will therefore yield valid recommendations for large numbers of clusters, or more generally for large numbers of degrees of freedom for variance estimation (Rao and Thomas 1988), as long as the covariance matrix V/n and hence the generalized design effects can be appropriately modelled. Since the Dirichlet mixture model used in this study yields generalized design effects having means and coefficients of variation that are typical of those found in practice, recommendations based on a large number of clusters or degrees of freedom (fifty or more) can be made with confidence. For small to moderate numbers of clusters, quantitative results may differ from design to design. Since the basic mechanisms underlying the results exhibited in this study, namely increasing non-normality of $\hat{\pi}$ for decreasing r plus the inherent conservativeness of Scheffé-like procedures, will apply in general, it is expected that the qualitative trends for the different statistics examined will be generalizable across a wide variety of designs, even when the number of clusters is not large. The basic aim of the study has been to identify procedures whose control of error rates is robust to variations in the study parameters, namely the number of categories, the number of clusters, the strength of clustering, and the skewness of the vector of category probabilities. The combination of parameters examined has covered much of the range likely to be encountered in practice, so it is reasonable to suggest that the robustness exhibited by the log and logit transformed Bonferroni intervals might extend to variations in survey design, for moderate numbers of clusters (or degrees of freedom). Further research on this question is clearly required.

Subject to these caveats, t -based Bonferroni simultaneous confidence intervals based on the logit transformation are recommended for assessing up to $k = 12$ proportions of varying magnitude, under realistic clustering conditions. If conservativeness is deemed to be an asset, the first-order modified Quesenberry-Hurst procedure can be safely used. Both procedures require only a knowledge of the variances (or design effects) of the estimated cell proportions.

ACKNOWLEDGMENTS

I wish to thank J.N.K. Rao for many valuable discussions and for his comments on a draft of this paper. Thanks are also due to Steve Brockwell both for useful suggestions and for excellent programming assistance, and to Paul Bertelman for programming assistance during the latter part of the project. Finally I wish to thank two reviewers and an associate editor for suggestions that significantly improved the presentation of the results. This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

REFERENCES

- ANDREWS, R.W., and BIRDSALL, W.C. (1988). Simultaneous confidence intervals: a comparison under complex sampling. Paper presented at the 1988 American Statistical Association Annual Meeting, Chicago.
- BAILEY, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformation of the cell frequencies. *Technometrics*, 22, 583-589.

- BLACK, D., and MYLES, J. (1986). Dependent industrialization and the Canadian class structure: a comparative analysis of Canada, the United States, and Sweden. *Canadian Review of Sociology and Anthropology*, 23, 157-181.
- BRIER, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67, 591-596.
- FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.
- FITZPATRICK, S., and SCOTT, A.J. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82, 875-878.
- GOLD, R.Z. (1963). Tests auxiliary to χ^2 tests in a Markov chain. *Annals of Mathematical Statistics*, 34, 56-74.
- GOODMAN, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.
- HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: Part I - simple goodness-of-fit, homogeneity and independence in a two-way table with applications to the Canada Health Survey (1978-1979). *Journal of Official Statistics*, 3, 117-132.
- HOCHBERG, Y., and TAMANE, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
- HOLT, D., SCOTT, A.J., and EWING, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Ser. A*, 143, 303-320.
- JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *The American Statistician*, 41, 335-337.
- MILLER, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Edition. New York: Springer-Verlag.
- QUESENBERY, C.P., and HURST, D.C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6, 191-195.
- RAO, J.N.K., and SCOTT, A.J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 58-66.
- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 261-230.
- RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.
- THOMAS, D.R. (1989). An investigation of simultaneous confidence interval procedures for proportions, under cluster sampling. Working Paper WPS 89-02, School of Business, Carleton University.
- THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Logistic Regression Under Complex Survey Designs

JORGE G. MOREL¹

ABSTRACT

Estimation procedures for obtaining consistent estimators of the parameters of a generalized logistic function and of its asymptotic covariance matrix under complex survey designs are presented. A correction in the Taylor estimator of the covariance matrix is made to produce a positive definite covariance matrix. The correction also reduces the small sample bias. The estimation procedure is first presented for cluster sampling and then extended to more complex situations. A Monte Carlo study is conducted to examine the small sample properties of F -tests constructed from alternative covariance matrices. The maximum likelihood estimation method where the survey design is completely ignored is compared with the usual Taylor's series expansion method and with the modified Taylor procedure.

KEY WORDS: Pseudo-likelihood; CPLX procedure; Cluster sampling; Adjusted covariance matrix.

1. INTRODUCTION

In the last few years a lot of attention has been given to the problems that arise when chi-square tests based on the multinomial distribution are applied to data obtained from complex sample designs. It has been shown that the effects of stratification and clustering on the chi-square tests may lead to a distortion of nominal significance levels. Holt, Scott and Ewings (1980) proposed modified Pearson chi-square statistics tests of goodness-of-fit, homogeneity, and independence in two-way contingency tables. Rao and Scott (1981) presented similar tests for complex sample surveys. In all these cases, the correction factor requires only the knowledge of variance estimates (or design effects) for individual cells. Bedrick (1983) derived a correction factor for testing the fit of hierarchical log linear models with closed form parameter estimates. Rao and Scott (1984) presented more extensive methods of using design effects to obtain chi-square tests for complex surveys. They generalized their previous results to multi-way tables. Fay (1985) presented the adjustments to the Pearson and likelihood test statistics through a jackknife approach.

The use of the conditional logistic model, Cox (1970), has become increasingly popular in the context of complex survey designs. Under suitable conditions, Binder (1983), proved the asymptotic normality of design-based sampling distribution for a family of parameter estimators that cannot be defined explicitly as a function of other statistics from the sample. His results are applied to binary logistic models. Further applications to the Canada Health Survey are also found in Binder *et al.* (1984).

Chambless and Boyle (1985) derived a general asymptotic distribution theory for stratified random samples with a fixed number of strata and increasing stratum sample sizes. Their theoretical results were illustrated with logistic regression and discrete proportional hazard-models. Albert and Lesaffre (1986) discussed the logistic discrimination method for classifying multivariate observations into one of several populations. They restrict their attention to discrimination between qualitatively distinct groups.

¹ Jorge G. Morel is Assistant Professor of the Department of Epidemiology and Biostatistics, University of South Florida, Tampa, Florida 33612.

Extensions to the case where the response consists of a polychotomous variable have been done by Bull and Pederson (1987) and Morel (1987). They show, by using Taylor's series expansion, that the large sample variance of the beta estimates has the form

$$H^{-1} G H^{-1}$$

where H^{-1} is the covariance matrix that wrongly results from assuming independence and multinomial distribution in the response vector, and G is a matrix whose estimation is based in the complex survey design.

More recently, Roberts, Rao and Kumar (1987) showed how to make adjustments that take into account the survey design in computing the standard chi-square and the likelihood ratio test statistics for logistic regression analysis involving a binary response variable. The adjustments are based on certain generalized design effects. Their results can be applied to cases where the whole population has been divided into I domains of study, a large sample is obtained for each domain, and in each domain a proportion π_i , $i = 1, 2, \dots, I$, is to be estimated. It is assumed

$$\pi_i = [1 + \exp(x_i \beta^0)]^{-1} \exp(x_i \beta^0), i = 1, 2, \dots, I,$$

where x_i is a k -vector of known constants derived from the i -th domain and β^0 is a k -vector of unknown parameters. This procedure may be most useful when only the summary table of counts and variance adjustment factors are available, instead of the complete data set.

In this paper an estimation procedure is presented for obtaining consistent estimators of the parameter vector of a generalized logistic model and its asymptotic covariance matrix when a complex sampling design is employed. The resulting estimated covariance matrix is always positive definite and asymptotically equivalent to the one obtained from Taylor's series expansion. A correction for reducing the small sample bias in the estimated covariance matrix is also introduced. It is shown, via a Monte Carlo study, that this correction levels off the inflated Type I error that arises from ignoring the complex survey, faster than the Taylor's series expansion. In this sense the correction proposed here produces, for small samples, results that are superior to the usual delta-method.

The new procedure will be termed, henceforth, the CPLX procedure, or simply CPLX. The maximum likelihood estimation method and the Taylor's series expansion method will be termed MLE and TAYLOR, respectively. The CPLX procedure has been incorporated into PC CARP, a personal computer program for variance estimation with large scale surveys, see Schnell *et al.* (1988).

2. LOGISTIC REGRESSION WITH CLUSTER SAMPLING

Consider first single-stage cluster sampling where n clusters or primary sampling units are taken with known probabilities with replacement from a finite population or without replacement from a very large population. Let m_j represent the size of the j -th cluster, $j = 1, 2, \dots, n$, and let $y_{j\ell}^*$, $\ell = 1, 2, \dots, m_j$ denote $(d + 1)$ dimensional classification vectors. The vector $y_{j\ell}^*$ consists entirely of zeros except for position r which will contain a one if the ℓ -th unit selected from the j -th cluster falls in the r -th category. Let $x_{j\ell}$ be a k -dimensional row vector of explanatory variables associated with the ℓ -th unit selected from the j -th cluster.

Then, for each $j = 1, 2, \dots, n$, and each $\ell = 1, 2, \dots, m_j$, the expectation of the r -th element of $y_{j\ell}^*$ is determined by a logistic relationship as

$$\begin{aligned}\pi_{j\ell r} &= E\{y_{j\ell r}\} = \left[1 + \sum_{s=1}^d \exp(x_{j\ell} \beta_s^0)\right]^{-1} \exp(x_{j\ell} \beta_r^0) \quad r = 1, 2, \dots, d \\ &= 1 - \sum_{s=1}^d \pi_{j\ell s}, \quad r = d + 1.\end{aligned}\quad (2.1)$$

Because the expected value function is nonlinear in the parameter vector $\beta^0 = (\beta_1^{0'}, \beta_2^{0'}, \dots, \beta_d^{0'})'$, it is necessary to use nonlinear estimation methods. Define the pseudo log-likelihood $L_n(\beta)$ as

$$L_n(\beta) = \sum_{j=1}^n \sum_{\ell=1}^{m_j} w_j (\log \pi_{j\ell}^*)' y_{j\ell}^*, \quad (2.2)$$

where $\pi_{j\ell}^* = (\pi_{j\ell 1}, \dots, \pi_{j\ell, d+1})'$ and w_j is the sampling weight for the $j\ell$ -th sampling unit. This function can be viewed as a weighted log likelihood function, where the weights are the sampling weights and the $y_{j\ell}^*$'s are distributed as multinomial random variables. If the sampling weights are all one, then (2.2) becomes the log-likelihood function under the assumption that the $y_{j\ell}^*$'s are independently multinomially distributed.

Let $\hat{\beta}_{\text{PSEUDO}}$ be the estimator of β^0 that maximizes (2.2). This estimator is a solution to the system of equations

$$\sum_{j=1}^n \sum_{\ell=1}^{m_j} w_j G(\beta, x_{j\ell}) [\text{Diag}(\pi_{j\ell}^*)]^{-1} (y_{j\ell}^* - \pi_{j\ell}^*) = \mathbf{0}, \quad (2.3)$$

where

$$G(\beta, x_{j\ell}) = [(\mathbf{I}_{d \times d}, \mathbf{0}_{d \times 1}) \otimes x_{j\ell}'] \Delta(\pi_{j\ell}^*),$$

$$\Delta(\pi_{j\ell}^*) = \text{Diag}(\pi_{j\ell}^*) - \pi_{j\ell}^* (\pi_{j\ell}^*)',$$

and \otimes denotes the Kronecker product.

The asymptotic normality of $\hat{\beta}_{\text{PSEUDO}}$ can be proved by defining the parameters of interest implicitly as in (2.2) and then by extending the results given in Binder (1983). An alternative approach can be derived by making use of the pseudo-likelihood assumption and Proposition 1 in Dale (1986). Binder and Dale both provide the necessary regularity conditions.

As n increases,

$$\sqrt{n}(\hat{\beta}_{\text{PSEUDO}} - \beta^0) = \sqrt{n}[H_n(\beta^0)]^{-1} U_n(\beta^0)$$

$$\xrightarrow{L} N_{dk}(\mathbf{0}, \lim_{n \rightarrow \infty} [H_n(\beta^0)]^{-1} G_n[H_n(\beta^0)]^{-1}) \quad (2.4)$$

where,

$$\begin{aligned}
 H_n(\underline{\beta}^0) &= \sum_{j=1}^n \sum_{\ell=1}^{m_j} w_j \Delta(\pi_{j\ell}) \otimes \mathbf{x}'_{j\ell} \mathbf{x}_{j\ell}, \\
 U_n(\underline{\beta}^0) &= \sum_{j=1}^n \sum_{\ell=1}^{m_j} w_j (y_{j\ell} - \pi_{j\ell}) \otimes \mathbf{x}'_{j\ell}, \\
 G_n &= \sum_{j=1}^n \sum_{\ell=1}^{m_j} w_j^2 \text{Var}(y_{j\ell}) \otimes \mathbf{x}'_{j\ell} \mathbf{x}_{j\ell},
 \end{aligned}$$

$y_{j\ell}$ and $\pi_{j\ell}$ are the vectors $\mathbf{y}_{j\ell}^*$ and $\pi_{j\ell}^*$, without their last elements, respectively and N_{dk} denotes a dk -multivariate normal distribution.

Nelder and Wedderburn (1972) have shown that under binomial assumption, the pseudo log-likelihood function (2.2) can be solved by an iterative weighted least-squares procedure. Haberman (1974, p.48) shows that under regularity conditions a modified Newton-Raphson converges to the maximum likelihood estimator for the multinomial case. His proof does not depend on the existence of any consistent estimator of $\underline{\beta}^0$ which allows the iterative algorithm to be initialized at $\hat{\underline{\beta}} = \mathbf{0}$. Jennrich and Moore (1975) proved that when the multinomial assumption holds, the common Gauss-Newton algorithm for finding the maximum likelihood estimator of $\underline{\beta}^0$ becomes the Newton-Raphson algorithm. Because of this equivalence of those algorithms and because a modified Newton-Raphson procedure always converge, we have adopted the modified Gauss-Newton algorithm described by Gallant (1987, p.318).

CPLX first finds $\hat{\underline{\beta}}_{\text{PSEUDO}}$ using an iterative procedure in which the estimate of $\underline{\beta}^0$ at the q -th step is

$$\begin{aligned}
 \hat{\underline{\beta}}_{[q, i(q)]} &= \hat{\underline{\beta}}_{[q-1, i(q-1)]} \\
 &+ (0.5)^{i(q)} [H_n(\hat{\underline{\beta}}_{[q-1, i(q-1)]})]^{-1} U_n(\hat{\underline{\beta}}_{[q-1, i(q-1)]})
 \end{aligned} \tag{2.5}$$

where $i(q)$ is a nonnegative integer such that

$$L_n(\hat{\underline{\beta}}_{[q, i(q)]}) > L_n(\hat{\underline{\beta}}_{[q-1, i(q-1)]}). \tag{2.6}$$

The modification of the iteration algorithm provided by $i(q)$ guarantees the convergence of the procedure. The iteration is initiated by setting $\hat{\underline{\beta}}_{(0)} = \mathbf{0}$. The algorithm is declared to have converged when the condition

$$\frac{L_n(\hat{\underline{\beta}}_{[q, i(q)]}) - L_n(\hat{\underline{\beta}}_{[q-1, i(q-1)]})}{|L_n(\hat{\underline{\beta}}_{[q, i(q)]})| + 10^{-5}} < \epsilon \tag{2.7}$$

is satisfied, where ϵ can be 10^{-8} .

Observe that a consistent estimator of $H_n(\underline{\beta}^0)$ is $H_n(\hat{\underline{\beta}}_{\text{PSEUDO}})$ and a distribution free estimator of G_n is

$$\mathbf{G}_n^* = (n-1)^{-1} n \sum_{j=1}^n (\mathbf{d}_j - \bar{\mathbf{d}}) (\mathbf{d}_j - \bar{\mathbf{d}})', \quad (2.8)$$

where

$$\mathbf{d}_j = \sum_{\ell=1}^{m_j} w_j (\mathbf{y}_{j\ell} - \pi_{j\ell}) \otimes \mathbf{x}'_{j\ell},$$

and $\bar{\mathbf{d}} = n^{-1} \sum_{j=1}^n \mathbf{d}_j$. If within each cluster, the $\mathbf{y}_{j\ell}^*$'s are independent and identically distributed according to a multinomial random vector with parameters $(\pi_{j\ell}^*, 1)$, then it can be easily shown that the expectation of \mathbf{G}_n is precisely $H_n(\underline{\beta}^0)$. In practice the $\pi_{j\ell}$'s in (2.8) are replaced with $\hat{\pi}_{j\ell}$ where $\hat{\pi}_{j\ell}$ is defined as in (2.1) with $\hat{\beta}_{\text{PSEUDO}}$ substituted by $\underline{\beta}^0$, and a small correction is applied to obtain the estimator

$$\hat{\mathbf{G}}_n = (n^* - k)^{-1} (n^* - 1) (n-1)^{-1} n \sum_{j=1}^n (\hat{\mathbf{d}}_j - \hat{\bar{\mathbf{d}}}) (\hat{\mathbf{d}}_j - \hat{\bar{\mathbf{d}}})', \quad (2.9)$$

where

$$\hat{\mathbf{d}}_j = \sum_{\ell=1}^{m_j} w_j (\mathbf{y}_{j\ell} - \hat{\pi}_{j\ell}) \otimes \mathbf{x}'_{j\ell},$$

$$\hat{\bar{\mathbf{d}}} = n^{-1} \sum_{j=1}^n \hat{\mathbf{d}}_j \quad \text{and} \quad n^* = \sum_{j=1}^n m_j.$$

The factor

$$(n^* - k)^{-1} (n^* - 1) (n-1)^{-1} n$$

reduces to $(n-k)^{-1}n$ if each cluster contains exactly one element. The factor $(n-k)^{-1}n$ is the degrees of freedom correction applied to the residual mean square for ordinary least squares in which k parameters are estimated. The quantity in (2.9) is well defined for two or more clusters and the factor $(n^* - k)^{-1} (n^* - 1)$ should reduce the small sample bias associated with using the estimated function to calculate deviations. Therefore, a consistent estimator of the asymptotic covariance matrix of $\hat{\underline{\beta}}_{\text{PSEUDO}}$ under the cluster sampling design is

$$\tilde{\mathbf{A}}_n = [\mathbf{H}_n(\hat{\underline{\beta}}_{\text{PSEUDO}})]^{-1} \hat{\mathbf{G}}_n [\mathbf{H}_n(\hat{\underline{\beta}}_{\text{PSEUDO}})]^{-1} \quad (2.10)$$

which can be used to test any hypothesis of the form $H_0: \mathbf{C} \underline{\beta}^0 = \underline{\delta}^*$. Under the null hypothesis, by Moore (1977)

$$(\mathbf{C} \hat{\underline{\beta}}_{\text{PSEUDO}} - \underline{\delta}^*)' [\mathbf{C} \tilde{\mathbf{A}}_n \mathbf{C}']^{-1} (\mathbf{C} \hat{\underline{\beta}}_{\text{PSEUDO}} - \underline{\delta}^*) \quad (2.11)$$

converges in law to a chi-square distribution with $\nu = \text{rank}(C \tilde{A}_n C')$ degrees of freedom. Here, $[C \tilde{A}_n C']^{-1}$ is any generalized inverse of $C \tilde{A}_n C'$.

The sums of squares and products matrix used in the construction of \hat{G}_n is based on n observations, where n is the number of clusters. By analogy to the Hotelling T^2 statistic, it is natural to adjust for degrees of freedom by multiplying (2.11) by the ratio

$$\frac{n - \nu}{\nu(n - 1)} \quad (2.12)$$

to obtain an approximate F statistic with ν and $n - \nu$ degrees of freedom. In our case, this adjustment has the disadvantage that ν may exceed n in a sample with a small number of clusters but a large number of individual elements.

The covariance matrix constructed as if the elemental observations are a simple random sample is biased, but it can be used to make a small sample adjustment in the estimated covariance matrix. One might view the usual small sample degrees-of-freedom adjustment as the operation of adding to an initial estimator of the covariance matrix the quantity $(n - \nu)^{-1} \nu \hat{V}$, where \hat{V} is also an estimator of the covariance matrix. In the usual case, \hat{V} is also the initial estimator. In our case, we make the adjustment using the covariance matrix based on the elements as the second \hat{V} . In our case, the use of the elemental covariance matrix has the advantage that the resulting sum is always positive definite. The adjustment is a function of the number of parameter estimated, dk . The adjustment is

(1) if $n > 3dk - 2$

$$\hat{A}_n = \tilde{A}_n + (n - dk)^{-1} (dk - 1) \gamma^* [H_n(\hat{\beta}_{\text{PSEUDO}})]^{-1}, \quad (2.13)$$

(2) if $n \leq 3dk - 2$

$$\hat{A}_n = \tilde{A}_n + 0.5 \gamma^* [H_n(\hat{\beta}_{\text{PSEUDO}})]^{-1}, \quad (2.14)$$

where $\gamma^* = \max(1, \text{tr}\{[H_n(\hat{\beta}_{\text{PSEUDO}})]^{-1} \hat{G}_n\} / dk)$. The upper bound of 0.5 for correction in (2.14) is arbitrary. Then, an approximate F -test with ν and $n - \nu$ degrees of freedom is obtained by substituting \hat{A}_n for \tilde{A}_n in (2.11) and dividing the resulting quadratic form by ν . In practice, the approximate degrees of freedom can be taken to be ν and infinity.

3. A MONTE CARLO STUDY

In this section a Monte Carlo study is conducted to examine properties of F -Tests (2.11) involving model parameters. Data are generated under two different sampling schemes that correspond to single-stage cluster sampling where the primary units all have the same sampling weight and are taken from an infinite population. In the first sampling scheme all the elements within the cluster have the same explanatory vector \mathbf{x} and therefore, the same conditional mean (2.1). This is the case where the logistic regression becomes weighted in the sense of several responses y 's with the same covariate vector \mathbf{x} . Different degrees of intra-class correlation are induced among the y 's belonging to the same cluster.

The second sampling scheme, unlike the first, places different vectors of covariates for different subjects within the cluster. The conditional mean (2.1) is also satisfied and different degrees of intra-class correlation are controlled. The effect of the intra-class correlation is studied for both sampling schemes under three different estimation procedures: MLE where the clustering effect is completely ignored, TAYLOR where the large sample covariance matrix (2.10) is used, and CPLX where the adjusted covariance matrix (2.13-2.14) is employed. These last two procedures, for large samples, are asymptotically equivalent. For small samples CPLX performs better than TAYLOR.

3.1 Sampling Scheme I

Suppose that x_1, x_2, \dots, x_n are k -dimensional independent and identically distributed normal random vectors with vector mean $\underline{\mu}$ and covariance matrix Σ . For each j , $j = 1, 2, \dots, n$, suppose that given x_j , the random vectors $y_{j0}^0, y_{j1}^0, \dots, y_{j,m_j}^0$ are independent and identically distributed multinomial random vectors, with parameters $(\pi_j^*, 1)$, where π_j^* satisfies the logistic function (2.1) evaluated at the true parameter vector $\underline{\beta}^0$ and at $x = x_j$. Let $U_{j1}, U_{j2}, \dots, U_{j,m_j}$ be a set of independent and identically distributed uniform (0,1) random variables. For a known and fixed ζ , $0 \leq \zeta \leq 1$, define

$$y_{j\ell}^* \equiv y_{j0}^0 \quad \text{if} \quad U_{j\ell} \leq \zeta \quad (3.1.1)$$

and

$$y_{j\ell}^* \equiv y_{j\ell}^0 \quad \text{if} \quad U_{j\ell} > \zeta, \quad (3.1.2)$$

$\ell = 1, 2, \dots, m_j$.

It can be shown that within the j -th cluster,

$$E(y_{j\ell}^*) = \pi_j^*, \quad (3.1.3)$$

$$\text{Cov}(y_{j\ell}^*, y_{jt}^*) = \Delta(\pi_j^*) \quad \text{if} \quad \ell = t, \quad (3.1.4)$$

and

$$\text{Cov}(y_{j\ell}^*, y_{jt}^*) = \zeta^2 \Delta(\pi_j^*) \quad \text{if} \quad \ell \neq t. \quad (3.1.5)$$

Therefore, given x_j , the random vector $t_j = \sum_{\ell=1}^{m_j} y_{j\ell}^*$ does not have a multinomial distribution. Instead

$$E(m_j^{-1} t_j) = \pi_j^* \quad (3.1.6)$$

and

$$\text{Var}(m_j^{-1} t_j) = [1 + \zeta^2 (m_j - 1)] m_j^{-1} \Delta(\pi_j^*), \quad (3.1.7)$$

where ζ^2 represents the intra-cluster correlation. Furthermore, if the m_j 's are constant, *i.e.*, $m_j = m$, the factor $\phi = [1 + \zeta^2(m - 1)]$ corresponds to the design effect defined by Kish (1965, p.258). An estimate of the design effect ϕ is

$$\hat{\phi} = (dk)^{-1} \left[\sum_{\ell=1}^{dk} \hat{a}_{(i,i)} / \hat{h}^{(i,i)} \right] \bar{w}^{-1}, \quad (3.1.8)$$

where $\hat{a}_{(i,i)}$ and $\hat{h}^{(i,i)}$ represent the (i,i) -th elements of \hat{A}_n in (2.13)-(2.14) and $[H_n(\hat{\beta}_{\text{PSEUDO}})]^{-1}$, respectively, and \bar{w} is the average of the sampling weights for the entire sample.

Under this sampling scheme, data $(x_j, y_{j\ell}^*)$, $j = 1, 2, \dots, n$, $\ell = 1, 2, \dots, m$, were generated with $k = 4$, $d = 3$, $m = 21$, and parameters

$$\underline{\mu} = (1, -2, 1, 5)', \quad (3.1.9)$$

$$\underline{\Sigma} = \text{Diag}(0, 25, 25, 25), \quad (3.1.10)$$

$$\underline{\beta}_1^0 = (-0.3, -0.1, 0.1, 0.2), \quad (3.1.11)$$

$$\underline{\beta}_2^0 = (0.2, -0.2, -0.2, 0.1), \quad (3.1.12)$$

and

$$\underline{\beta}_3^0 = (-0.1, 0.3, -0.3, 0.1). \quad (3.1.13)$$

Based on (3.1.9)–(3.1.13), 1000 sets of samples with n clusters of size m , were generated according to (3.1.1)–(3.1.2) for different values of n , ζ^2 , and ϕ . The estimated Type I errors obtained from comparing the F -tests of $H_0: \underline{\beta} = \underline{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ were computed under the three different estimation procedures: MLE, CPLX and TAYLOR. A measure of the distortion of the estimated Type I errors relative to the nominal 0.05 is the relative bias which is defined as

$$(0.05)^{-1} | \text{Estimated Type I error} - 0.05 |. \quad (3.1.14)$$

Relative biases of the estimated Type I errors are reported in Table 3.1. For data generated with no intra-class correlation, ($\zeta^2 = 0$) the MLE procedure, as it is expected, provides small relative bias of the estimated nominal 5% level. CPLX produces in this case relative biases slightly greater than MLE. This is the penalty of estimating extra parameters in (2.13-2.14).

The MLE procedure shows a strong distortion of the estimated Type I error when a positive intra-class correlation is present. This distortion increases as the intra-class correlation ζ^2 gets bigger. In the case where $\zeta^2 = 0.15$ ($\phi = 4$) the relative bias of the estimated Type I error is about 18 indicating an inflated Type I error of about 95%. For the CPLX procedure, the

Table 3.1
Relative Bias of the Estimated Type I Error for the F -test of $H_0: \beta = \beta^0$
with nominal 0.05 Level under Sampling Scheme I

n	ξ^2	ϕ	Procedure		
			MLE	CPLX	TAYLOR
20	0.00	1	0.24	0.60	16.42
20	0.05	2	9.66	3.68	17.06
20	0.10	3	15.24	3.98	17.44
20	0.15	4	17.74	4.00	17.70
30	0.00	1	0.08	0.06	12.82
30	0.05	2	9.84	1.20	13.74
30	0.10	3	15.52	1.76	14.22
30	0.15	4	17.74	1.86	14.68
40	0.00	1	0.04	0.32	9.66
40	0.05	2	9.98	0.82	9.62
40	0.10	3	16.20	1.02	11.66
40	0.15	4	17.74	1.80	11.66
50	0.00	1	0.06	0.50	7.40
50	0.05	2	9.76	1.44	8.38
50	0.10	3	16.00	1.96	9.32
50	0.15	4	17.80	2.20	9.70
100	0.00	1	0.06	0.90	2.68
100	0.05	2	10.02	1.66	3.90
100	0.10	3	16.26	2.06	4.70
100	0.15	4	17.78	2.24	5.10
200	0.00	1	0.02	0.74	1.28
200	0.05	2	10.46	1.00	1.64
200	0.10	3	16.30	0.88	1.88
200	0.15	4	18.00	1.52	2.12
400	0.00	1	0.02	0.44	0.70
400	0.05	2	10.14	0.66	0.90
400	0.10	3	16.56	0.64	1.00
400	0.15	4	17.86	0.56	0.84
800	0.00	1	0.08	0.32	0.40
800	0.05	2	10.36	0.22	0.36
800	0.10	3	16.04	0.68	0.80
800	0.15	4	18.12	0.50	0.54

relative bias decreases as the sample size increases from $n = 20$ to the cutting point of correction (2.14) which is 34 in this case. Then it slightly increases as the sample size approaches $n = 100$ and then decreases as the sample size keeps getting bigger. This pattern will be observed throughout the whole simulation. It represents the effect of the correction (2.13-2.14) in small samples.

The Taylor procedure has large relative biases when the sample sizes are small. It varies from 17 to 7 for sample sizes between $n = 20$ and $n = 50$. For large samples both methods CPLX and TAYLOR, provide as expected, similar results. In general, the CPLX shows relative biases smaller than the TAYLOR method.

If the F statistics used for testing $H_0: \underline{\beta} = \underline{\beta}^0$ are multiplied by the number of parameters being tested, the resulting statistic is distributed as a chi-square random variable with 12 degrees of freedom. The Monte Carlo means and variances for these chi-square statistics are presented in Table 3.2.

As expected, the MLE method produces means and variances around 12 and 24, respectively, when the design effect ϕ is one. CPLX has in this case means around 12 with greater variances that decrease when the sample size gets bigger. However, in the presence of any intra-class correlation, the means and variances under MLE are too large, while CPLX shows consistency with the asymptotic theory and the correction introduced in (2.13-2.14). The TAYLOR method has extremely high variances when the sample size is small. A possible explanation for this is that in some replications of the simulation the covariance matrix (2.10) was ill-conditioned producing very large quadratic forms for (2.11). This problem attenuates when the sample size is bigger. Both methods, CPLX and TAYLOR, become asymptotic equivalent for large samples.

Monte Carlo properties for the estimator (3.1.8) of the design effect are presented in Table 3.3 for both CPLX and TAYLOR methods. The CPLX procedure shows smaller biases and slightly large standard errors. Both methods perform fairly well.

For each category $r, r = 1, 2, 3$ and each covariate $s, s = 1, 2, 3, 4$, “ t ” statistics for the individual coefficient estimates were also computed as

$$“t” = [\text{Var}(\hat{\underline{\beta}}_{rs})]^{-0.5}(\hat{\underline{\beta}}_{rs} - \underline{\beta}_{rs}^0). \tag{3.1.15}$$

The twelve “ t ” statistics provided by the CPLX estimation procedure were grouped together and the simulated percentiles were computed. Similar computations were performed for the MLE “ t ” statistics. Consequently, for each run the percentiles are based on 12,000 “ t ” values. Once these percentiles were calculated, the relative biases were estimated as

$$(\text{Standard Normal Percentile})^{-1} | \text{Estimated Percentile} - \text{Standard Normal Percentile} |. \tag{3.1.16}$$

The results of the relative bias for the estimated 5th and 95th percentiles for the “ t ” statistics are presented in Table 3.4 for both MLE and CPLX procedures. Under the MLE it is expected that these relative biases be close to $\phi^{0.5} - 1$. This is true because the “ t ” statistics under MLE are inflated by the factor $\phi^{0.5}$. This is clearly seen in Table 3.4 under the two columns for the MLE percentiles. The CPLX procedure has satisfactory relative biases for small sample. These biases become negligible, as expected, when the sample sizes get bigger.

Table 3.2
Monte Carlo Properties of the Chi-square Statistic of $H_0: \beta = \beta^0$
under Sampling Scheme I

n	ξ^2	ϕ	Procedure					
			MLE		CPLX		TAYLOR	
			Mean	Variance	Mean	Variance	Mean	Variance
20	0.00	1	11.5	22.2	12.0	32.7	81.9	12x10 ³
20	0.05	2	23.9	134.3	16.5	81.2	116.6	8x10 ⁴
20	0.10	3	34.2	239.9	16.6	77.8	94.5	12x10 ³
20	0.15	4	43.8	403.2	17.3	89.3	140.3	19x10 ⁴
30	0.00	1	11.8	25.1	11.2	28.5	35.1	702.3
30	0.05	2	23.8	121.4	13.2	41.2	34.1	691.6
30	0.10	3	35.8	268.1	13.8	46.3	41.2	12x10 ²
30	0.15	4	46.7	450.1	14.1	51.1	44.5	16x10 ²
40	0.00	1	12.2	24.3	11.9	30.3	25.8	268.3
40	0.05	2	23.2	96.5	12.6	33.6	25.4	201.4
40	0.10	3	35.4	247.7	13.5	43.3	29.1	340.4
40	0.15	4	46.2	428.9	13.8	44.4	30.2	331.4
50	0.00	1	11.9	25.5	12.4	34.6	21.0	140.8
50	0.05	2	23.9	112.5	13.7	43.8	22.7	153.6
50	0.10	3	35.8	231.0	14.3	46.0	24.6	195.8
50	0.15	4	46.7	424.0	14.5	55.4	25.2	234.6
100	0.00	1	12.1	23.6	13.2	35.0	15.8	55.0
100	0.05	2	23.9	102.6	13.8	39.2	16.5	62.1
100	0.10	3	36.5	233.9	14.6	47.0	17.6	75.8
100	0.15	4	47.5	350.4	14.6	43.0	17.9	70.6
200	0.00	1	11.7	24.1	12.6	32.4	13.6	38.2
200	0.05	2	23.9	93.9	13.1	33.1	14.1	39.1
200	0.10	3	35.7	194.1	13.3	31.5	14.3	37.4
200	0.15	4	48.0	399.6	13.5	35.7	14.6	42.7
400	0.00	1	11.9	24.9	12.3	29.3	12.7	31.3
400	0.05	2	24.1	96.6	12.7	29.2	13.1	31.3
400	0.10	3	36.9	208.5	13.1	29.2	13.6	31.4
400	0.15	4	47.3	390.7	12.7	31.6	13.1	34.0
800	0.00	1	11.9	24.0	12.1	26.4	12.3	27.2
800	0.05	2	24.0	99.3	12.3	27.3	12.5	28.2
800	0.10	3	36.4	239.3	12.6	30.1	12.8	31.1
800	0.15	4	48.7	396.3	12.6	26.7	12.7	27.5

Table 3.3
Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme I

<i>n</i>	ζ^2	ϕ	Procedure			
			CPLX		TAYLOR	
			Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.00	1	0.28	0.23	0.23	0.22
20	0.05	2	0.01	0.63	0.35	0.48
20	0.10	3	0.07	0.93	0.40	0.70
20	0.15	4	0.15	1.15	0.46	0.85
30	0.00	1	0.33	0.22	0.17	0.20
30	0.05	2	0.14	0.62	0.25	0.47
30	0.10	3	0.08	0.88	0.30	0.66
30	0.15	4	0.04	1.18	0.33	0.90
40	0.00	1	0.26	0.18	0.14	0.18
40	0.05	2	0.14	0.53	0.19	0.42
40	0.10	3	0.10	0.83	0.22	0.67
40	0.15	4	0.07	1.13	0.25	0.91
50	0.00	1	0.18	0.18	0.11	0.17
50	0.05	2	0.09	0.48	0.16	0.41
50	0.10	3	0.07	0.75	0.18	0.64
50	0.15	4	0.04	0.97	0.21	0.83
100	0.00	1	0.07	0.13	0.06	0.13
100	0.05	2	0.04	0.34	0.08	0.32
100	0.10	3	0.01	0.54	0.10	0.51
100	0.15	4	0.01	0.69	0.11	0.65
200	0.00	1	0.03	0.10	0.03	0.09
200	0.05	2	0.02	0.25	0.04	0.24
200	0.10	3	0.01	0.38	0.05	0.36
200	0.15	4	0.01	0.49	0.05	0.48
400	0.00	1	0.01	0.07	0.01	0.07
400	0.05	2	0.01	0.19	0.02	0.19
400	0.10	3	0.00	0.27	0.02	0.27
400	0.15	4	0.00	0.37	0.02	0.37
800	0.00	1	0.01	0.05	0.01	0.05
800	0.05	2	0.00	0.13	0.01	0.13
800	0.10	3	0.00	0.19	0.01	0.18
800	0.15	4	0.00	0.24	0.01	0.24

Table 3.4
Relative Bias of the Estimated 5th and 95th Percentiles for the “t” Statistics
for the Coefficient Estimates under Sampling Scheme I

n	ξ^2	$\phi^{0.5} - 1$	Procedure			
			MLE Percentile		CPLX Percentile	
			5th	95th	5th	95th
20	0.00	0.00	0.02	0.00	0.10	0.09
20	0.05	0.41	0.40	0.38	0.04	0.02
20	0.10	0.73	0.68	0.65	0.07	0.04
20	0.15	1.00	0.84	0.79	0.07	0.04
30	0.00	0.00	0.00	0.02	0.10	0.09
30	0.05	0.41	0.43	0.38	0.01	0.02
30	0.10	0.73	0.73	0.70	0.02	0.01
30	0.15	1.00	0.97	0.91	0.01	0.01
40	0.00	0.00	0.01	0.01	0.07	0.08
40	0.05	0.41	0.38	0.41	0.03	0.02
40	0.10	0.73	0.70	0.72	0.03	0.01
40	0.15	1.00	0.96	0.93	0.01	0.03
50	0.00	0.00	0.01	0.01	0.05	0.07
50	0.05	0.41	0.43	0.40	0.00	0.01
50	0.10	0.73	0.71	0.70	0.01	0.00
50	0.15	1.00	0.97	0.96	0.02	0.01
100	0.00	0.00	0.00	0.02	0.01	0.00
100	0.05	0.41	0.42	0.42	0.02	0.01
100	0.10	0.73	0.71	0.74	0.01	0.03
100	0.15	1.00	1.03	0.99	0.04	0.04
200	0.00	0.00	0.01	0.01	0.00	0.00
200	0.05	0.41	0.42	0.43	0.01	0.01
200	0.10	0.73	0.71	0.72	0.01	0.01
200	0.15	1.00	1.00	1.00	0.02	0.02
400	0.00	0.00	0.01	0.01	0.01	0.01
400	0.05	0.41	0.39	0.40	0.01	0.00
400	0.10	0.73	0.76	0.77	0.03	0.04
400	0.15	1.00	1.02	0.89	0.02	0.00
800	0.00	0.00	0.00	0.01	0.00	0.01
800	0.05	0.41	0.43	0.44	0.01	0.02
800	0.10	0.73	0.76	0.70	0.02	0.01
800	0.15	1.00	1.07	1.04	0.04	0.02

3.2 Sampling Scheme II

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be a set of k -dimensional independent and identically distributed normal random vectors with vector mean $\underline{\mu}$ and covariance matrix $\underline{\Sigma}_B$. These vectors \mathbf{x} represent cluster means for the explanatory variables in the logistic function (2.1). Suppose that for the j -th cluster, $j = 1, 2, \dots, n$, $\mathbf{x}_{j0}^0, \mathbf{x}_{j1}^0, \dots, \mathbf{x}_{j,m_j}^0$ are independent and identically distributed normal random vectors with vector mean \mathbf{x}_j^0 and covariance matrix $\underline{\Sigma}_W$. Given $\mathbf{x}_{j\ell}^0$, $\ell = 0, 1, \dots, m_j$, the $(d + 1)$ -dimensional random vector $\mathbf{y}_{j\ell}^0$ has a multinomial distribution with parameters $(\pi_{j\ell}^0, 1)$, where the elements of $\pi_{j\ell}^0$ satisfy the logistic function (2.1) evaluated at the true parameter vector $\underline{\beta}^0$ and at $\mathbf{x} = \mathbf{x}_{j\ell}^0$. Furthermore, suppose that given the $\mathbf{x}_{j\ell}^0$'s, the $\mathbf{y}_{j\ell}^0$'s are independent.

Let $U_{j1}, U_{j2}, \dots, U_{j,m_j}$ be m_j independent and identically distributed uniform (0,1) random variables that are also jointly independent from the $\mathbf{x}_{j\ell}^0$'s and from the $\mathbf{y}_{j\ell}^0$'s. Let ζ be a fixed and known number, $0 \leq \zeta \leq 1$. Then define $(\mathbf{x}_{j\ell}, \mathbf{y}_{j\ell}^*)$, $\ell = 1, 2, \dots, m_j$ in the following way:

$$(\mathbf{x}_{j\ell}, \mathbf{y}_{j\ell}^*) \equiv (\mathbf{x}_{j0}^0, \mathbf{y}_{j0}^0) \text{ if } U_{j\ell} \leq \zeta \quad (3.2.1)$$

and

$$(\mathbf{x}_{j\ell}, \mathbf{y}_{j\ell}^*) \equiv (\mathbf{x}_{j\ell}^0, \mathbf{y}_{j\ell}^0) \text{ if } U_{j\ell} > \zeta. \quad (3.2.2)$$

Observe that within each cluster, the $\mathbf{x}_{j\ell}$'s all have the same vector of conditional means \mathbf{x}_j and that the covariance matrix between $\mathbf{x}_{j\ell}$ and \mathbf{x}_{jt} is $\underline{\Sigma}_W$ if $\ell = t$ and $\zeta^2 \underline{\Sigma}_W$ otherwise. Also, note that the conditional mean of each $\mathbf{y}_{j\ell}^*$ is the logistic function (2.1) evaluated at $\underline{\beta}^0$ and $\mathbf{x} = \mathbf{x}_{j\ell}$, and that the vectors $(\mathbf{x}_{j\ell}, \mathbf{y}_{j\ell}^*)$, $\ell = 1, 2, \dots, m_j$, exhibit an intra-class correlation of ζ^2 and an approximate design effect of $\phi = [1 + \zeta^2 (m - 1)]$ when all the m_j 's are constant.

Data $(\mathbf{x}_{j\ell}, \mathbf{y}_{j\ell}^*)$, $j = 1, 2, \dots, n$, $\ell = 1, 2, \dots, m_j$, were generated under this cluster sampling scheme with $k=4$, $d=3$, and parameters

$$\underline{\mu} = (1, -6, 4, 8)', \quad (3.2.3)$$

$$\underline{\Sigma}_B = \text{Diag}(0, 25, 25, 49), \quad (3.2.4)$$

$$\underline{\Sigma}_W = \text{Diag}(0, 25, 36, 36), \quad (3.2.5)$$

$$\underline{\beta}_1^0 = (0.30, -0.05, -0.06, 0.08), \quad (3.2.6)$$

$$\underline{\beta}_2^0 = (0.06, -0.08, -0.10, 0.07), \quad (3.2.7)$$

and

$$\underline{\beta}_3^0 = (0.70, -0.08, -0.10, 0.11), \quad (3.2.8)$$

Based on (3.2.3)–(3.2.8), 1000 sets of samples with n clusters of size $m_j = m = 6$, were generated according to (3.2.1)–(3.2.2) for different values of n , ζ^2 and ϕ . The relative biases defined in (3.1.14) of the estimated Type I errors from comparing the F -tests of $H_0: \underline{\beta} = \underline{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ are presented in Table 3.5 under three different estimation techniques: MLE, CPLX and TAYLOR.

In the presence of intra-class correlation, there is a strong distortion of the Type I error for MLE even in the case where ζ^2 is relatively small ($\zeta^2 = 0.2$) for cluster size $m = 6$. This distortion is reflected in the relative bias which ranges from approximately 7 to 18. These values indicate inflated Type I errors between 40% and 95%. The CPLX procedure provides satisfactory relative biases even for the case of small samples. The TAYLOR procedure has too high values for small samples. It becomes equivalent to CPLX for large samples. One more time CPLX seems to be superior to TAYLOR when the sample size is small.

Table 3.5
Relative Bias of the Estimated Type I Error for the F -test of $H_0: \underline{\beta} = \underline{\beta}^0$
with Nominal 0.05 Level under Sampling Scheme II

n	ζ^2	ϕ	Procedure		
			MLE	CPLX	TAYLOR
20	0.0	1	0.54	0.46	13.52
20	0.2	2	7.30	0.46	12.96
20	0.4	3	13.70	0.68	13.96
20	0.6	4	17.08	0.60	14.72
30	0.0	1	0.28	0.78	7.78
30	0.2	2	8.72	0.72	8.16
30	0.4	3	14.84	0.72	9.32
30	0.6	4	17.50	0.82	9.23
40	0.0	1	0.36	0.56	5.16
40	0.2	2	9.28	0.56	5.76
40	0.4	3	15.38	0.64	5.84
40	0.6	4	17.76	0.70	5.80
50	0.0	1	0.44	0.56	3.44
50	0.2	2	9.34	0.08	4.86
50	0.4	3	15.48	0.38	4.36
50	0.6	4	17.56	0.46	4.16
100	0.0	1	0.16	0.04	1.26
100	0.2	2	9.46	0.26	1.46
100	0.4	3	15.94	0.44	2.00
100	0.6	4	18.16	0.14	1.46
200	0.0	1	0.10	0.26	0.76
200	0.2	2	10.20	0.34	0.82
200	0.4	3	16.22	0.02	0.48
200	0.6	4	18.06	0.06	0.52

Table 3.6
Monte Carlo Properties of the Chi-square Statistic of $H_0: \underline{\beta} = \underline{\beta}^0$
under Sampling Scheme II

n	ζ^2	f	Procedure					
			MLE		CPLX		TAYLOR	
			Mean	Variance	Mean	Variance	Mean	Variance
20	0.0	1	11.3	18.9	10.2	19.7	40.5	15x10 ²
20	0.2	2	20.3	62.8	10.5	21.4	39.2	11x10 ²
20	0.4	3	28.3	106.4	10.5	18.4	111.3	42x10 ⁵
20	0.6	4	35.2	152.6	10.3	18.2	11x10 ³	50x10 ⁹
30	0.0	1	11.6	21.6	9.4	16.3	22.0	147.3
30	0.2	2	21.8	75.2	9.9	17.5	22.7	161.2
30	0.4	3	30.4	117.6	9.8	16.5	24.3	224.6
30	0.6	4	39.3	191.0	9.5	14.5	24x10 ²	60x10 ⁸
40	0.0	1	11.6	21.3	9.9	19.4	18.1	86.7
40	0.2	2	22.4	76.5	10.4	18.3	18.9	80.8
40	0.4	3	31.8	153.2	10.2	17.8	19.2	90.4
40	0.6	4	41.4	223.1	10.1	16.9	19.3	104.4
50	0.0	1	11.5	19.9	10.6	20.0	16.1	56.9
50	0.2	2	22.7	80.6	11.4	23.9	17.5	70.9
50	0.4	3	32.3	160.1	11.1	22.9	17.4	73.7
50	0.6	4	41.7	262.3	10.7	19.7	17.0	63.8
100	0.0	1	11.8	21.5	11.8	25.2	13.9	36.2
100	0.2	2	22.9	87.3	11.9	27.0	14.0	38.5
100	0.4	3	34.7	191.8	12.3	27.9	14.4	40.7
100	0.6	4	45.1	297.7	12.0	25.0	14.1	37.2
200	0.0	1	12.0	23.8	12.1	26.3	13.0	30.3
200	0.2	2	24.0	88.6	12.4	25.9	13.3	30.0
200	0.4	3	34.5	175.2	12.0	23.3	12.8	27.0
200	0.6	4	46.8	320.0	12.2	24.0	13.0	27.9

Monte Carlo properties of the chi-square statistics of $H_0: \underline{\beta} = \underline{\beta}^0$ (chi-square = 12 × F) are presented in Table 3.6 for the three estimation procedures under study. CPLX shows means and variances slightly below 12 and 24, respectively, when the sample sizes are small. This underestimation vanishes when the sample size increases. The TAYLOR procedure has too large means and variances when the sample size is small. For instance, for $\zeta^2 = 0.6$, the variance is in the order of billions when n is 30 or less. For large samples, both CPLX and TAYLOR, seem to provide similar results. The MLE method has acceptable results only when $\zeta^2 = 0.00$. Otherwise the estimated mean and variances are too large.

Table 3.7
Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme II

n	ξ^2	ϕ	Procedure			
			CPLX		TAYLOR	
			Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.0	1	0.48	0.22	0.04	0.20
20	0.2	2	0.16	0.53	0.26	0.42
20	0.4	3	0.05	0.87	0.34	0.72
20	0.6	4	0.01	1.24	0.39	1.03
30	0.0	1	0.49	0.18	0.02	0.16
30	0.2	2	0.25	0.48	0.19	0.40
30	0.4	3	0.19	0.84	0.24	0.69
30	0.6	4	0.16	1.12	0.27	0.94
40	0.0	1	0.38	0.16	0.02	0.14
40	0.2	2	0.22	0.45	0.14	0.38
40	0.4	3	0.16	0.70	0.20	0.60
40	0.6	4	0.16	0.98	0.19	0.86
50	0.0	1	0.27	0.14	0.02	0.13
50	0.2	2	0.15	0.42	0.12	0.37
50	0.4	3	0.12	0.67	0.15	0.60
50	0.6	4	0.11	0.89	0.16	0.81
100	0.0	1	0.12	0.10	0.01	0.10
100	0.2	2	0.06	0.32	0.07	0.31
100	0.4	3	0.05	0.50	0.07	0.48
100	0.6	4	0.06	0.59	0.07	0.57
200	0.0	1	0.05	0.07	0.01	0.07
200	0.2	2	0.03	0.24	0.03	0.23
200	0.4	3	0.02	0.34	0.04	0.33
200	0.6	4	0.02	0.40	0.03	0.40

Monte Carlo properties for the estimator of the design effect proposed in (3.1.8) are presented in Table 3.7 under the CPLX and TAYLOR procedures. The TAYLOR procedure seems to perform slightly better than CPLX for small samples. Both procedures, in general, provide reasonable values. They seem to be equivalent for large samples.

Table 3.8
Relative Bias of the Estimated 5th and 95th Percentiles for the “*t*” Statistics
for the Coefficient Estimates under Sampling Scheme II

<i>n</i>	ζ^2	$\phi^{0.5} - 1$	Procedure			
			MLE Percentile		CPLX Percentile	
			5th	95th	5th	95th
20	0.0	0.00	0.01	0.00	0.15	0.18
20	0.2	0.41	0.37	0.32	0.06	0.09
20	0.4	0.73	0.63	0.57	0.02	0.05
20	0.6	1.00	0.79	0.74	0.05	0.05
30	0.0	0.00	0.02	0.00	0.15	0.16
30	0.2	0.41	0.39	0.38	0.10	0.10
30	0.4	0.73	0.68	0.63	0.07	0.08
30	0.6	1.00	0.91	0.86	0.05	0.07
40	0.0	0.00	0.01	0.00	0.12	0.15
40	0.2	0.41	0.39	0.40	0.10	0.06
40	0.4	0.73	0.65	0.60	0.07	0.09
40	0.6	1.00	0.99	0.89	0.04	0.05
50	0.0	0.00	0.01	0.01	0.10	0.10
50	0.2	0.41	0.39	0.40	0.05	0.04
50	0.4	0.73	0.73	0.72	0.02	0.01
50	0.6	1.00	1.00	0.95	0.00	0.01
100	0.0	0.00	0.01	0.01	0.04	0.05
100	0.2	0.41	0.40	0.37	0.02	0.02
100	0.4	0.73	0.72	0.73	0.00	0.00
100	0.6	1.00	1.00	1.02	0.01	0.02
200	0.0	0.00	0.02	0.01	0.00	0.01
200	0.2	0.41	0.40	0.45	0.01	0.02
200	0.4	0.73	0.71	0.68	0.01	0.01
200	0.6	1.00	1.03	0.95	0.02	0.02

The relative biases (3.1.16) of the 5th and 95th percentiles of the “*t*” statistics (3.1.15) are presented in Table 3.8 under the MLE and CPLX procedures. MLE has a relative bias, as expected, close to zero in the absence of intra-class correlation. This bias increases when the ζ^2 gets bigger. On the other hand, CPLX has small relative bias in general and for large sample this bias becomes negligible.

4. EXTENSION TO STRATIFIED SAMPLING AND MORE COMPLEX DESIGNS

A generalization of CPLX procedure to stratified sampling can be done as follows. Suppose that the population has been divided into $i = 1, 2, \dots, L$ strata. Let m_{ij} represent the size of the j -th cluster in the i -th stratum, n_i the number of clusters selected in the i -th stratum, and $y_{ij\ell}^*$ the multinomial response of the ℓ -th element in the j -th cluster in the i -th stratum, $\ell = 1, 2, \dots, m_{ij}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, L$. It is assumed that $\pi_{ij\ell}^*$, the expected value of $y_{ij\ell}^*$, satisfies the logistic relationship (2.1) for a given explanatory vector $x_{ij\ell}$.

A consistent estimator of β^0 , say $\hat{\beta}_{\text{PSEUDO}}$, can be found by maximizing the function

$$L_n(\beta) = \sum_{i=1}^L \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij} (\log \pi_{ij\ell}^*)' y_{ij\ell}^*. \tag{4.1}$$

Algorithm (2.5) is performed with three indexes i, j, ℓ . The adjustment given by (2.13) and (2.14) is applied with

$$n = \sum_{i=1}^L n_i, \tag{4.2}$$

$$H_n(\hat{\beta}_{\text{PSEUDO}}) = \sum_{i=1}^L \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij} \Delta(\hat{\pi}_{ij\ell}^*) \otimes x'_{ij\ell} x_{ij\ell}, \tag{4.3}$$

$$\hat{G} = [(n^* - k)^{-1} (n^* - 1)] \sum_{i=1}^L (n_i - 1)^{-1} n_i (1 - f_i) \sum_{j=1}^{n_i} (\hat{d}_{ij} - \hat{a}_i)(\hat{d}_{ij} - \hat{a}_i)', \tag{4.4}$$

$$\hat{d}_{ij} = \sum_{\ell=1}^{m_{ij}} w_{ij} (y_{ij\ell} - \hat{\pi}_{ij\ell}) \otimes x'_{ij\ell}, \tag{4.5}$$

$$\hat{a}_i = n_i^{-1} \sum_{j=1}^{n_i} \hat{d}_{ij}, \tag{4.6}$$

f_i = sampling rate of i -th stratum, and (4.7)

$$n^* = \sum_{i=1}^L \sum_{j=1}^{n_i} m_{ij}. \tag{4.8}$$

The estimation procedure can be extended in a stepwise manner to multi-stage sampling designs by maximizing (4.1) up to elemental units. The summation of (4.3) should be extended in order to include all the final sampling units. The key part is (4.4). The construction of \hat{G} must be based on the complex survey. This could be a difficult task for multi-stage sampling. Results for stratified two-stage sampling are presented in Fuller, *et al.* (1986, p. 82).

5. SUMMARY

In this paper, we have outlined a methodology for obtaining asymptotic normal estimators of the parameters of a generalized logistic function involving a multinomial response variable under complex survey designs. A consistent estimator of the asymptotic covariance matrix under the complex sampling design is (2.10), which results from the usual Taylor's series expansion. This covariance matrix produces for large samples correct Type I errors for the F -tests involving model parameters. More important, it is shown that correction (2.13-2.14) provides a covariance matrix that reduces the small sample bias. This adjusted covariance matrix has some important characteristics:

1. It levels off the inflated Type I error, originated from ignoring the complex survey, faster than the usual delta-method.
2. It is positive definite when $H_n(\hat{\beta}_{\text{PSEUDO}})$ is positive definite regardless if (2.9) is singular or not.
3. It is asymptotic equivalent to (2.10).

The results of a Monte Carlo study were reported in Section 3. Data satisfying the logistic conditional mean (2.1) were generated under two different single-stage cluster sampling schemes. It was studied, among other things, the effect of the intra-class correlation and the design effect on the relative biases of the estimated Type I errors for the F -tests of $H_0: \beta = \beta^0$. The simulation showed, as expected, a strong relative bias when the naive maximum likelihood method is employed. For small samples, the Monte Carlo results favor the use of the adjusted covariance matrix over the one that arises from the usual delta-method.

ACKNOWLEDGEMENTS

This work was begun while the author was a student at Iowa State University. I thank Professor Wayne A. Fuller for introducing me to the topic and for suggesting a number of the small sample modifications that were incorporated into the estimation procedure. The author wants also to thank the referees for useful comments.

REFERENCES

- ALBERT, A., and LESAFFRE, E. (1986). Multiple group logistic discrimination. *Computers and Mathematics with Applications*, 12A, 209-224.
- BEDRICK, E.J. (1983). Adjusted chi-square tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.
- BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BINDER, D.A., GRATTON, M.A., HIDIROGLOU, M.A., KUMAR, S., and RAO, J.N.K. (1984). Analysis of categorical data from surveys with complex designs: some Canadian experiences. *Survey Methodology*, 10, 141-156.
- BULL, S.B., and PEDERSON, L.L. (1987). Variance for polychotomous logistic regression using complex survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

- CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, Theory and Methods*, 14, 1377-1392.
- COX, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.
- DALE, J.R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society, Ser. B*, 48, 48-59.
- FAY, R.E. (1985). A jackknife chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). *PC CARP*. Statistical Laboratory, Iowa State University, Ames, Iowa.
- GALLANT, A.R. (1987). *Nonlinear Statistical Methods*. New York: John Wiley & Sons.
- HABERMAN, S.J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.
- HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Ser. A*, 143, 303-320.
- JENNRICH, R.I., and MOORE, R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Section on Statistical Computing, American Statistical Association*.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- MOORE, D.S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72, 131-137.
- MOREL, J. (1987). Multivariate nonlinear models for vectors of proportions: A generalized least squares approach. Unpublished Ph.D. dissertation. Iowa State University, Ames, Iowa.
- NELDER, J.A., and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Ser. A*, 135, 370-384.
- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.

Randomized Response Sampling from Dichotomous Populations with Continuous Randomization

LeROY A. FRANKLIN¹

ABSTRACT

A randomized response model for sampling from dichotomous populations is developed in this paper. The model permits the use of continuous randomization and multiple trials per respondent. The special case of randomization with normal distributions is considered, and a computer simulation of such a sampling procedure is presented as an initial exploration into the effects such a scheme has on the amount of information in the sample. A portable electronic device is discussed which would implement the presented model. The results of a study taken, using the electronic randomizing device, is presented. The results show that randomized response sampling is a superior technique to direct questioning for at least some sensitive questions.

KEY WORDS: Randomized response; Randomization with continuous distributions; Computer simulation.

1. INTRODUCTION

Surveys often seek to estimate the proportion of individuals satisfying a particular condition. If the condition involves a highly personal or controversial subject (*e.g.*, seeking new employment, sexual behavior) or of an illegal nature (*e.g.* drug usage, criminal activities), survey respondents may be reluctant to answer honestly or may refuse to answer a direct question as to whether they satisfy the condition of interest. In such cases, it is difficult to make inferences about proportions on the basis of a survey in which sensitive questions are asked directly.

Randomized response sampling plans utilize a stochastic or randomizing device to enable respondents to provide answers to sensitive questions without fully revealing information regarding the sensitive issue. The actual outcome of the device for a particular respondent is observed by the respondent but not by the interviewer. However, the properties of the device are known to the experimenter, and this enables the experimenter to make inferences about the proportion of interest without knowing specifically about any single individual. The stochastic device introduces noise into the information-gathering process, but the resulting loss of information may be preferable to the uncontrollable noise introduced by nonresponse or lying when direct questions are used.

The original randomized response model was proposed by Warner (1965) and involved a dichotomous randomization for a dichotomous population. His model was studied from a Bayesian viewpoint in Winkler and Franklin (1979). The randomized response model with two or more trials per respondent was introduced by Gould, Shah and Abernathy (1969) and further developed by Liu and Chow (1976). Both papers demonstrated the superiority of the multiple trials per respondent in improving the efficiency of the estimate over the single trial model of Warner's. However, both also note that multiple trials might produce simultaneously

¹ Dr. LeRoy A. Franklin, Department of System and Decision Sciences, Indiana State University, School of Business, Terre Haute, Indiana 47809.

growing suspicion and lowered “truth telling” over the single trial model. The survey paper prepared by Horvitz, Greenberg, and Abernathy (1976) discusses several other plans with discrete randomization devices. In addition a thorough theoretical development and review of results is contained in the recent volume by Chaudhuri and Mukerjee (1988) entitled “Randomized Response: Theory and Techniques.” A more general model, using either discrete or continuous randomization, is presented in Warner (1971) and these more general models were discussed from a Bayesian viewpoint by Pitz (1980), Smouse (1984), and O’Hagen (1987). A few surveys have actually been undertaken, some showing the randomized response methods are superior to direct survey methods (*e.g.* Gould *et al.* 1969 and Liu and Chow 1976) and a few others of uncertain results (*e.g.* Brewer 1981). However, only Poole (1974) developed a specific continuous randomization distribution (uniform) to estimate a continuous distribution and this was implemented by having respondents report their answer multiplied by a number chosen randomly from a random number table.

In this paper, we consider a randomized response model for sampling from a dichotomous population, but using a continuous randomization distribution. With Warner’s original randomized response technique, the randomizing device determines which question the respondent answers. But with the method developed in this paper, the question for a respondent is fixed by whether or not he belongs to the sensitive group. The randomization here chooses values from two distributions (one for “yes” and the other for “no”) and the respondent provides the value appropriate to his group membership. Multiple trials are incorporated into the model by having the respondent provide a single multi-digit response. This provides a potential benefit over usual multiple trial techniques in that the respondent perceives he/she has provided just one answer when in fact the multi-digit response incorporates several trials of the respondent.

The general model, for which the randomization can be handled via any type of distribution, is presented in Section 2. The special case in which the randomization involves normal distributions is discussed in Section 3, along with an approximating procedure for assessing the effect of randomization and multiple trials per respondent. Section 4 presents a computer simulation investigating the role that specific choices of means and standard deviations play in the efficiency of surveying by using normal distribution randomization with multiple trials. Section 5 presents a way of implementing normal distributions as the randomizing distribution through the use of a computerized, electronic device that generates and displays random normal values. Such a device was felt to be potentially superior to “drawing cards” or “flipping a spinner” since these methods may not be properly implemented by the respondent or the interviewer. The results of a survey taken using that electronic device to investigate five sensitive questions are examined in Section 6. Finally, a summary and a brief discussion of design issues are considered in Section 7.

2. THE MODEL

Suppose that we are interested in θ , the proportion of individuals belonging to Group A among the members of a particular population. A simple random sample of n individuals is chosen from the population with $n \geq 1$, where we assume that the population is large enough relative to n so that the sampling process can be viewed effectively as sampling with replacement. A total of k trials are conducted with each respondent, where $k \geq 1$. On trial j for respondent i , random values are drawn from the distribution functions G_{ij} and H_{ij} . The respondent sees both values and is asked to report the value from G_{ij} if he or she belongs to Group A and

the value from H_{ij} otherwise. The researcher knows the exact form of G_{ij} and H_{ij} but sees only the value reported by the respondent, denoted by z_{ij} , and, thus, does not know from which distribution it came.

Inferences must be made about θ based on the kn sample observations z_{ij} , with $i = 1, \dots, n$ and $j = 1, \dots, k$. For convenience, we assume in the remainder of this paper that G_{ij} and H_{ij} are absolutely continuous with corresponding densities g_{ij} and h_{ij} ; the development for the discrete case is analogous. The conditional density function of z_{ij} given θ is $\theta g_{ij}(z_{ij}) + (1 - \theta) h_{ij}(z_{ij})$, and the likelihood function for the entire experiment is:

$$L(z | \theta) = \prod_{i=1}^n \left[\theta \prod_{j=1}^k g_{ij}(z_{ij}) + (1 - \theta) \prod_{j=1}^k h_{ij}(z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1, \tag{2.1}$$

where $z = (z_1, \dots, z_n)$ and $z_i = (z_{i1}, \dots, z_{ik})$.

Expanding the likelihood function using the binomial theorem allows the likelihood function to be written in the form

$$L(z | \theta) = \sum_{t=0}^n \alpha_t \theta^t (1 - \theta)^{n-t} \text{ where } 0 \leq \theta \leq 1 \text{ and} \tag{2.2}$$

$$\alpha_t = \sum_{s=1}^c \left[\prod_{i \in C_{ts}} \prod_{j=1}^k g_{ij}(z_{ij}) \right] \left[\prod_{i \notin C_{ts}} \prod_{j=1}^k h_{ij}(z_{ij}) \right], \text{ with} \tag{2.3}$$

C_{t1}, \dots, C_{tc} representing the $c = \binom{n}{t}$ combinations of t items out of n . Here $\theta^t (1 - \theta)^{n-t}$ is the Bernoulli likelihood conditional upon exactly t respondents being in Group A, and α_t is the likelihood of z given t . The mixture form in 2.2 arises because we are unable to observe a specific t in our sample.

A special case of (2.1) arises when we assume that the same randomizing distributions are used for all n respondents. Thus, $g_{ij} = g_j$ and $h_{ij} = h_j$ for $i = 1 \dots n$ and thus (2.1) reduces to

$$L(z | \theta) = \prod_{i=1}^n \left[\theta \prod_{j=1}^k g_j(z_{ij}) + (1 - \theta) \prod_{j=1}^k h_j(z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1. \tag{2.4}$$

Whichever the form, in order to find the maximum likelihood estimates, a direct computer grid search must be made. This is feasible since θ is only a one-dimensional quantity and is restricted to the interval from 0 to 1. This can be easily accomplished by using well-known search techniques applied to the log of the likelihood function. (See, for example, Kennedy and Gentle 1980).

3. RANDOMIZATION WITH NORMAL DISTRIBUTIONS

Although any continuous distribution (e.g. Weibull, uniform, etc.) can be used as the randomizing distribution in the model discussed in Section 2, in this section only the normal distribution will be examined. Furthermore, suppose that the same randomization distributions are used for all respondents, so that form (2.4) is the appropriate likelihood. Thus, g_j and h_j are normal densities with means μ_{gj} and μ_{hj} and standard deviations σ_{gj} and σ_{hj} , respectively. Then the likelihood function in Section 2 can be related to these normal densities.

The amount of information that can be obtained about θ obviously depends on the means and standard deviations that are chosen. At one extreme, if $\mu_{gj} = \mu_{hj}$ and $\sigma_{gj} = \sigma_{hj}$ for $j = 1, \dots, k$, then θ drops out of the likelihood function and \underline{z} (the sample) will provide no information about θ . At the other extreme, if $|\mu_{gj} - \mu_{hj}| \rightarrow \infty$ for any j with σ_{gj} and σ_{hj} fixed or if $\sigma_{gj} \rightarrow 0$ and $\sigma_{hj} \rightarrow 0$ for any j with a fixed $|\mu_{gj} - \mu_{hj}| \neq 0$, then we are effectively able to determine which group each respondent belongs to and the sampling process thus approaches Bernoulli sampling in θ .

An approximation to $L(\underline{z} | \theta)$ as developed by Winkler and Franklin (1979) makes it easier to assess the effect of randomization and multiple trials with the choice of specific means and standard deviations. That is, for each sample, we can approximate the actual likelihood function given by (2.4) with an approximate likelihood function of the form

$$L^*(r^*, n^* | \theta) = \theta^{r^*} (1 - \theta)^{n^* - r^*}. \quad (3.1)$$

Taking the first and second derivations of the log of the approximating likelihood (3.1) and solving to find the maximum ($\hat{\theta}$) and the curvature at that maximum yields:

$$\hat{\theta} = \frac{r^*}{n^*} \quad (3.2)$$

and
$$\left[\frac{\partial^2 \log L^*(r^*, n^* | \theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} = - \frac{n^*}{\hat{\theta} (1 - \hat{\theta})}. \quad (3.3)$$

Next taking the first derivative of the log of the exact likelihood (2.4) and setting it to equal zero gives the equation that will yield the exact maximum likelihood estimate for θ :

$$\sum_{i=1}^n \frac{\gamma_i - \eta_i}{\theta \gamma_i + (1 - \theta) \eta_i} = 0 \text{ where } \gamma_i = \prod_{j=1}^k g_j(z_{ij}), \eta_i = \prod_{j=1}^k h_j(z_{ij}). \quad (3.4)$$

A grid search produces for (3.4) its solution ($\hat{\theta}_r$). Taking the second derivative of the log of the exact likelihood (2.4) yields:

$$\left[\frac{\partial^2 \log L(\underline{z} | \theta)}{\partial \theta^2} \right] = - \sum_{i=1}^n \frac{[\gamma_i - \eta_i]^2}{[\theta \gamma_i + (1 - \theta) \eta_i]^2}. \quad (3.5)$$

Substituting $\hat{\theta}_r$ into (3.5) gives the curvature of the actual log likelihood at $\hat{\theta}_r$ (the maximum). Equations (3.2) and (3.3) are two equations in two unknowns, r^* and n^* . Setting (3.2) = $\hat{\theta}_r$ and (3.3) = (3.5) allows us to solve for r^* and n^* so that the approximating log likelihood has the same maximum $\hat{\theta} = \hat{\theta}_r$, and curvature at that maximum as does the actual log likelihood. Thus, the randomized response sample outcome of \underline{z} can be thought of as approximately equivalent to a non-randomized response sample (*i.e.* regular Bernoulli sampling) with r^* members out of n^* in the sensitive group. In this sense, n^* can be thought of as a rough measure of the amount of information in the randomized response sample which is of size n .

4. A COMPUTER SIMULATED INVESTIGATION
OF THE CHOICE OF MEANS AND
STANDARD DEVIATIONS

To investigate the impact of a given set of means and standard deviations for the normal randomizing distributions as well as the impact the size of θ and k (the number of trials) has upon r^* and n^* the randomized response sampling process was simulated by generating, via computer, repeated samples from a Bernoulli process with parameter θ and k sets of two-digit responses for each sample. In our simulation, we let $\mu_{gj} = 50$, $\mu_{hj} = 40$, and $\sigma_{gj} = \sigma_{hj} = \sigma$ for $j = 1, \dots, k$. We considered two values of θ (.10 and .25), two values of σ (6 and 9), three values of n (50, 200, and 500), and three values of k (1, 2, and 3). Such values were chosen since they will register two-digit deviates that would overlap in distribution considerably and provided then a bench mark for later choices in the actual survey environment. For each of the 36 combinations of parameters, we replicated the sampling procedure 25 times. The solutions of r^* and n^* were found numerically for each sample, and the average values of n^* for the 25 replications with each set of parameter values are given in Table 1.

The average values of n^* vary considerably. At the worst extreme, when $\sigma = 9$, $\theta = .10$, and only one trial per respondent is used, n^* tends to be only 10-15 percent of n . On the other hand, when $\sigma = 6$, $\theta = .25$, and three trials are used per respondent, n^* is about 75 percent of n . As expected, the average value of n^* (the effective sample size) increases as n (the number of respondents) increases or as k (the number of trials per respondent) increases. In addition, decreasing σ or increasing θ also leads to a higher n^* .

For each combination of parameters, the mean and variance of $\hat{\theta}$ over the 25 trials were determined. The average values of $\hat{\theta}$ are very close (within 5%) to the corresponding values of θ , and the variance of $\hat{\theta}$ tends to increase as the average n^* decreases and, hence, tends to validate the simulation.

Table 1
Average Values of the Effective Sample Size (n^*) for Various Sample Sizes (n) and the
Number of Trials per Respondent (k)

<i>n</i>	<i>k</i>	$\theta = .10$		$\theta = .25$	
		$\sigma = 6$	$\sigma = 9$	$\sigma = 6$	$\sigma = 9$
50	1	16.2	7.0	17.3	9.2
	2	27.3	13.1	30.6	17.8
	3	32.6	18.1	38.2	23.6
200	1	58.3	24.8	79.0	41.2
	2	103.1	49.6	124.4	72.9
	3	136.6	77.7	151.0	97.7
500	1	148.4	59.6	196.9	103.6
	2	261.1	129.3	309.5	181.2
	3	345.8	193.1	375.6	242.7

5. A PORTABLE, COMPUTERIZED RANDOMIZING DEVICE

Randomized-response sampling, using randomization with normal distributions and multiple trials, provides flexibility to the experimenter, who can select means and variances as well as the number of respondents and the number of trials per respondent. However, this flexibility is not of any value, unless the sampling scheme actually can be implemented in practice. The sampling scheme utilizing Bernoulli randomization can be implemented in a number of ways (*e.g.*, with cards or colored beads). However, the scheme developed in this paper requires generation of random normal values by some portable device.

A computerized, electronic device was built around the Intel 8080 microprocessor to generate and display random normal values. Each value is obtained by summing 16 uniformly distributed random numbers and transforming that sum to achieve a normal deviate with the desired mean and standard deviation. From the Central Limit Theorem, the resulting values should be approximately normally distributed, and extensive tests indicate that the values produced by the device do indeed behave like random normal values. This technique was chosen over other possible methods of generating normal deviates due to the simplicity of programming such a method in machine instructions for this specific microprocessor. For more details concerning the generation of the random normal values and the testing of the device, see Franklin (1977), Kennedy and Gentle (1980), as well as Knuth (1969).

The final, resulting device was approximately the size of a cigar box and is easily held in the hand. Power can be supplied either by a battery pack or by an extension cord.

For display purposes, the random normal values are truncated to two digits, and the device is designed to display six such two-digit numbers simultaneously in "windows" of six digits each. One window displays values chosen from g_1 , g_2 , and g_3 which appears as a single six-digit number in the "Yes" window. The other window displays values chosen from h_1 , h_2 , and h_3 which also appears as a single six-digit number for "No". The six means and standard deviations are stored permanently in the device, but they can be changed easily by using a small, detachable keyboard.

The actual surveying process is accomplished in the following manner. First, the interviewer asks the respondent a sensitive question about Group A. The respondent then pushes a button to activate the device, and two six-digit numbers appear in the windows within about one quarter of a second. If the respondent is a member of Group A, the number in the first window (the "Yes" window) is reported; otherwise, the number in the second window (the "No" window) is reported. To convince the respondent of the "randomness" of the values, he or she is encouraged to press the button several times and to observe the resulting numbers before the sensitive question is actually asked. Note that although $k = 3$, the respondent perceives a response as a single six-digit number, and we are thus actually obtaining three trials with a single six digit response. Hence, the advantage of multiple trials per respondent is exploited without the usual accompanying disadvantages coming into play.

6. SURVEY RESULTS AND CONCLUSIONS

Two simultaneous, but independent, surveys were conducted on the campus of a large urban university of students enrolled in that university. The first asked five sensitive questions of a respondent by the direct question method. The second asked the same five sensitive questions of a different respondent but using Randomized Response Sampling with continuous randomization implemented by the electronic device presented in the previous section. For the

study $k = 3$ and $\mu_{g_1} = \mu_{g_2} = \mu_{g_3} = 40$ and $\mu_{h_1} = \mu_{h_2} = \mu_{h_3} = 50$ with $\sigma_{g_j} = \sigma_{h_j} = 5$ for $j = 1, 2, 3$. These values were chosen in accordance to the finding of the computer simulation discussed in Section 4. A different group of students was systematically selected (one in five) for each of the two surveys from students on the campus and individually interviewed. Each student surveyed was given a brief introduction as to the purpose of the survey and asked if they wished to participate. Less than 10% of all individuals stopped by both survey teams declined to participate. If the individual was willing to participate, he/she was then asked to provide his/her social security number to verify that he/she was, indeed, enrolled in the university. All respondents of both surveys had their social security number checked against an administrative master list of students and those not recorded as enrolled students were eliminated from the study (less than 5 percent of those surveyed).

Requiring their social security number also deliberately injected the element of associating the individual's identity with his responses. For many surveys (*i.e.* telephone, mail-in questionnaires, house-to-house surveys, *etc.*), this is the case and plays a significant role in the willingness of a respondent to answer truthfully. It was felt that it was precisely in such "revealing" circumstances that randomized response sampling can benefit the researcher most. The resulting sample sizes for the direct and randomized response methods were $n_1 = 473$ and $n_2 = 477$. The five sensitive questions were:

Q1 — "Have you ever cheated on an exam here at this university?"

Q2 — "Would you ever cheat on your income tax?"

Q3 — "Would you ever steal from an employer?"

Q4 — "Have you smoked any marijuana in the last 30 days?"

Q5 — "Have you ever participated in a homosexual act?"

All five questions were felt to be sufficiently sensitive so that any gains by randomized response sampling over direct sampling could be easily apparent. In addition, as a final question, the respondents in the randomized response group were asked "Do you think your friends would be more willing to tell the truth if they were asked sensitive questions by this technique?" This was asked in an effort to measure the acceptance and confidence of the person being interviewed that this particular randomized response technique did provide personal protection and anonymity.

The estimates of the proportion of respondents who are in the sensitive group are presented in Table 2 for both direct ($\hat{\theta}_{id}$) and randomized response ($\hat{\theta}_{ir}$) for question i along with the estimate of n_i^* (the effective sample size) for the randomized response method using the method discussed in Section 3. Also is presented the z value of a one-sided test of hypothesis $H_0: \theta_{id} - \theta_{ir} = 0$ vs $H_a: \theta_{id} - \theta_{ir} < 0$, along with the observed p -values. The tests were conducted using n_1 and n_i^* as sample sizes and hence give a much more conservative result than if n_1 and n_2 were utilized.

It is noteworthy that the randomized response method gave a higher estimate of θ for each of the five sensitive questions than the direct survey method. Furthermore, for Questions 1, 2, and 5, the randomization response method gave statistically significantly higher estimates of θ (p -values $< .001$ for all three) than the direct survey method. Hence, there seems to be conclusive evidence that, at least for some sensitive issues, the randomized response method with continuous randomization does provide better estimates of population proportions. It should also be noted that by our choices of μ_{g_j} , μ_{h_j} , σ_{g_j} and σ_{h_j} and $k = 3$ that n_i^* typically was 75 to 85 percent of the original sample size n_2 and thus most of the information was "recovered" by our randomized response method.

Table 2

Estimates of θ and Results of Testing Equality of θ 's for Direct and Randomized Response Sampling with Respective Sample Sizes of $n_1 = 473$ and $n_2 = 477$

Question <i>i</i>	$\hat{\theta}_{id}$	$\hat{\theta}_{ir}$	Effective sample size		<i>p</i> -value
			n_1^*	z-value	
1	.0634	.2013	394.5	6.098	< .0001
2	.1797	.2941	408.1	3.997	< .0001
3	.1078	.1207	384.8	.583	.2810
4	.1882	.1942	409.5	.234	.4091
5	.0042	.0355	339.0	3.341	.0004

Furthermore, it is instructive to consider the nonsignificant results for Questions 3 and 4. This information (if the three significant results are ignored) could lead an observer to conclude that randomized response techniques are not particularly advantageous over direct questioning. However, in the light of the three significant differences revealed, this lack of significance perhaps could be interpreted as the question really was not "sensitive enough" to lead to dramatic differences in θ 's or even that the question was "so sensitive" that the respondent chose to lie even with the randomized response technique. In addition, Question 1 "Have you ever cheated on an exam?" seemed to the experimenter to be relatively "unsensitive" but in retrospect the answer to this question when tied to the social security number of the respondent (given before the questioning process started) presented a much more threatening circumstance than was initially realized. Thus, perhaps some of the confusion about the efficacy of the randomized response technique is related to the "true sensitivity" of the question for the interviewee as opposed to the "perceived sensitivity" by the interviewer or experimenter. These aspects need further examination.

Finally, 88.9% (424 of the 477) felt "their friends would be more likely to answer truthfully sensitive questions by this randomized response technique." While some reservations may be expressed by the respondents' "desire to please the interviewer," nevertheless, this overwhelming percentage coupled with the significant differences already discussed seem strong evidence that this technique was accepted and felt to be protective of the interviewee.

7. DISCUSSION

The model developed in this paper permits the use of continuous, as well as discrete, randomizing distributions in utilizing randomized response sampling from a dichotomous population. In order to implement the model using randomization with normal distributions, a computerized, electronic device was also developed and discussed. The device is portable, has programmable means and standard deviations for the six normal distributions and provides from a single six digit response, three separate two digit trials. Such a system has both potential advantages and disadvantages over other randomized response techniques.

First, as alluded to in the introduction, a computerized randomizing device could be superior to the standard randomized response methods of "drawing cards" or "flipping a spinner" since these methods may not be properly implemented by either the respondent or the interviewer which would induce uncontrolled error. (See Abernathy, Greenberg and Horvitz (1970)

for a discussion of the problems of “insufficient card shuffling” and “card loss” as well as insufficient interviewer training). Since the production of the randomizing values is computerized, the distributional problems that can and have accompanied the use of cards, beads, and spinners are eliminated because the problem of “random selection of values” is taken out of the hands of the interviewer *and* respondent and placed in the “hands” of the computer. If the computerized device fails, it is usually a complete, catastrophic crash of the whole chip which is readily apparent and very, very rare.

The second (and perhaps greatest) advantage is in the ability of the device to present a choice of two numbers each six digits in length from which the respondent chooses to answer “yes” or “no”. But what seems to the respondent as a single six digit answer is in fact three separate two digit answers and in effect provides three trials per respondent. Thus, the benefits of multiple trials per respondent are gained but, since the respondent is unaware of the multiple trials format, without the usual accompanying disadvantages (noted by Liu and Chow 1976) coming into play.

In addition, the freedom to choose the six means and six standard deviations provides the experimenter with additional flexibility over standard randomized response techniques. For instance, if it is felt that the differences in the first two digits are most noticeable to respondents, the experimenter can make μ_{h_1} and σ_{h_1} close to (or even equal to) μ_{g_1} , and σ_{g_1} , respectively. Similarly, if the middle two digits might receive the least attention, the experimenter could attempt to gain the most information from these values by separating μ_{h_2} and μ_{g_2} the furthest. It is also possible to wire the displays in other than the obvious manner. For instance, the two digits of the first random normal value could appear as the fifth and second digits of the six digit number instead of the first and second digits. This flexibility in wiring, together with the choices of parameters should provide a sampling scheme that is quite informative to the researcher without seemingly to threaten the respondent.

It should also be noted that while for this particular microprocessor it was convenient to utilize randomization with normal distributions, several other continuous distributions (*e.g.* uniform, Weibull) or even multi-valued discrete distributions (*e.g.* multinomial or poisson) could have been used. Further investigation into newer microprocessors as well as different randomizing distributions is recommended.

There are, however, some potential disadvantages associated with this particular randomized response technique. The cost of such a randomizing device since it involves a microprocessor is the order of fifteen hundred to two thousand dollars to produce. However, its versatility in wiring and programming would hopefully allow a device to be used in many investigations over several years and thus help to defray its rather high cost.

More difficult to quantify is the respondent’s perception of the computerized device and the degree of confidence or suspicion he/she might have about the device. Do respondents fear that the computerized device is somehow “storing” their answer that somehow later can be deciphered to expose them? From the survey results, it seems that greater truth telling was secured by using the computerized randomizing devices over the direct survey method. Nevertheless, further study is recommended to compare this randomized response technique which uses the computerized device with other more standard randomized response techniques.

In practice, several matters are relevant in the consideration of design issues (*i.e.*, the selection of means and standard deviations for the device). In order to gain more information for a given sample size, we should increase $|\mu_{g_j} - \mu_{h_j}|$ and decrease σ_{g_j} and σ_{h_j} for $j = 1, 2, 3$. However, as this is done, it will become clearer to the respondent that, despite the randomization, the response is very revealing concerning the respondent’s group membership. As a result, the respondent may not answer honestly or may refuse to answer. Additional study is needed

to determine optimal values for choice of means and standard deviations. The results in Table 1 give some indication of the effects of varying a common standard deviation. But from a practical viewpoint, the field survey seemed to indicate that the choice of means separated by two standard deviations was able to both gain the confidence of the respondent and (with the multiple trials) to gain back from 75 to 85 percent of the original sample size without the usual "loss of confidence" that accompanies multiple trial techniques.

In particular, the field trial compared the direct survey techniques with the randomized response using the electronic device discussed with $\mu_{h_j} = 40$ and $\mu_{g_j} = 50$ and $\sigma_{h_j} = \sigma_{g_j} = 5$ for $j = 1, 2, 3$ for the normal, randomizing distributions. Of the five sensitive questions which were asked of the two (independent) groups, the randomized response method provided significantly greater estimates ($p < .001$) than the direct method for three of the questions. In addition, 88.9% of the subjects interviewed by the randomized response technique felt "their friends would be more likely to tell the truth if they were asked sensitive questions by this technique". Thus, it seems that (for at least certain questions), this randomized response sampling technique achieved greater honesty in response than the direct sampling method.

The question of protection of the respondent's privacy needs to be discussed. It is not ethical to tell the respondent that his or her group membership is disguised by the randomization, if, in fact, the disguise is transparent to the researcher (e.g. for example, by recording only even numbers for "YES" and only odd numbers for "NO"). With the electronic device that has been discussed, it seems indeed possible to provide true privacy without losing much information. If the means and standard deviations are programmed into the device and are not provided to an interviewer, the interviewer will find it very difficult to discriminate between group members and non-group members in the interviewing process, particularly if the wiring is "scrambled". Thus, the flexibility that enables us to gain information without threatening the respondent also helps to disguise the actual group membership from the interviewer.

ACKNOWLEDGEMENTS

The author wishes to thank the referees and an Associate Editor for their constructive comments.

REFERENCES

- ABERNATHY, J.R., GREENBERG, B.G., and HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.
- BARNARD, G.A. (1976). Discussion on the invited and contributed papers. *International Statistical Review*, 44, 226.
- BREWER, K.R.W. (1981). Estimating marijuana usage using randomized response some paradoxical findings. *Australian Journal of Statistics*, 23, 139-148.
- CAMPBELL, C., and JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27, 229-231.
- CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker, Inc..
- CHOW, L.P., LIU, P.T., and MOSELY, W.H. (1973). A new randomized response technique for study of contemporary social problems. Presented at the 101st Annual Meeting of the American Public Health Association, Statistics Section.

- FRANKLIN, L.A. (1977). A Bayesian approach to randomized response sampling. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.
- GOULD, A.L., SHAH, B.U., and ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Section on Social Statistics American Statistical Association*, 351-359.
- HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistics Review*, 44, 181-196.
- KENNEDY, W.J., and GENTLE, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker, Inc..
- KNUTH, D.E. (1969). *Semi Numerical Algorithms*, (Volume 2). New York: Addison Wesley.
- LIU, P.T., and CHOW, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.
- O'HAGAN, A. (1987). Bayes linear estimates for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.
- PITZ, G.F. (1980). Bayesian analysis of randomized response models. *Psychological Bulletin*, 87, 209-212.
- POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.
- SMOUSE, E.P. (1984). A note on Bayesian least squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- WINKLER, R.L., and FRANKLIN, L.A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.

Small Area Estimates of Proportions Via Empirical Bayes Techniques

BRENDA MacGIBBON¹ and THOMAS J. TOMBERLIN²

ABSTRACT

Empirical Bayes techniques are applied to the problem of "small area" estimation of proportions. Such methods have been previously used to advantage in a variety of situations, as described, for example, by Morris (1983). The basic idea here consists of incorporating random effects and nested random effects into models which reflect the complex structure of a multi-stage sample design, as was originally proposed by Dempster and Tomberlin (1980). Estimates of proportions can be obtained, together with associated estimates of uncertainty. These techniques are applied to simulated data in a Monte Carlo study which compares several available techniques for small area estimation.

KEY WORDS: Logistic regression; Random effects models; Bayes estimation; EM algorithm.

1. INTRODUCTION

1.1 The Problem

Complex multi-stage surveys are used to obtain estimates of proportions in many research disciplines (*e.g.*, epidemiology, economics, criminology *etc.*). Not only are estimates for local areas and other special subgroups required, but there is also a need for reliable measures of the accuracy of these estimates. This suggests to us the need for improved methodologies for this estimation problem and related statistical inference.

In addition, the techniques based on the standard normal theory used by Fay and Herriot (1979) to estimate income, a continuous random variable, in small areas are no longer directly applicable to the problem of estimating proportions for discrete outcome variables. Here, it is the logit transform of the proportion, not the proportion itself, that will be modelled in a linear way. This creates the same problems of estimation as in classical statistical logistic regression theory. (See Haberman 1978.) Unfortunately, fewer attempts have been made to solve this obviously more complex problem in small area estimation.

In order to address the problem of inference from a relatively thinly spread complex, multi-stage survey to small areas or domains not necessarily included in the survey, we have chosen an explicitly model-based approach. This was proposed originally by Dempster and Tomberlin (1980) for the estimation of census undercount from a post-enumeration survey. The methodology uses both a random effects, multiple logistic regression model and empirical Bayes techniques. This directly yields estimates of uncertainty associated with the estimated proportions for small areas via a Bayesian paradigm. This explicitly model-based method differs substantially from the implicitly model-based approach of the synthetic estimation techniques of Gonzalez and Hoza (1976, 1978), Gonzalez and Waksberg (1975), and others.

¹ Brenda MacGibbon, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8 and Département de mathématiques et d'informatique, Université du Québec à Montréal, C.P. 8888, Suc. "A", Montréal, Québec H3C 3P8.

² Thomas J. Tomberlin, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8.

As a typical complex survey will often be a nested structure of primary sampling units (PSU's), secondary sampling units (SSU's) within PSU's, tertiary sampling units (TSU's) within SSU's and, finally, households within TSU's; the explicitly model-based approach will allow us to take into account the complexity of the sample design. The purpose of introducing a random effects model is to allow the data to determine, by empirical Bayes techniques, an appropriate compromise between the classical unbiased estimates which depend only on data in the specific local area, and the fixed effects estimates which pool information across areas.

In Section 1.2, a literature review is given and a solution to the problem of estimating proportions for small areas is proposed. The model and its associated estimates are made explicit in Sections 2 and 3 respectively. The results are applied to simulated data in a Monte Carlo study presented in Section 4.

1.2 The Review and a Proposed Solution to the Problem

Because of the growing need for small area statistics in recent years, and because reliable estimates for small areas or subdomains are not usually directly available by classical sample survey methods, several researchers have focused on the problem of small area estimation. This has necessitated the use of explicitly or implicitly model-based methods which allow for "borrowing strength" across small areas in order to increase the effective sample size for estimation, and hence the accuracy of the resulting estimates. Although much of the research in this area has applied linear model techniques and concentrated on the estimation of means or totals, rather than proportions, a discussion of the literature on these estimators and the criteria used to evaluate them can add valuable insight into our problem.

Classical theory dictates that estimators should be design-consistent and, if possible essentially design-unbiased. However these estimators are not always particularly useful when the sample sizes are small.

Gonzalez (1973) described the method of synthetic estimation as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." It seems its first reported use was by the U.S. National Center for Health Statistics (1968) for the calculation of state estimates of long and short term disability rates. Various authors subsequently tried to formalize this concept of synthetic estimation, in particular, for means of continuous outcome variables, using both *ad hoc* and model-based approaches. Gonzalez (1973), Gonzalez and Waksberg (1975), Gonzalez and Hoza (1976) and Levy and French (1978) used previous census data to form post-strata which are subsequently used to combine information across small areas under the assumption that the mean response is similar across a section of these areas. Levy (1971), Ericksen (1973, 1974) and O'Hare (1976) employed regression methods in order to incorporate auxiliary information in small area estimation. The accuracy of this method has been evaluated in terms of its average sampling mean squared error over all small areas in a region.

Ericksen (1974) warned that there is no systematic methodology for the assessment of the bias or accuracy of synthetic estimators. Despite these shortcomings, synthetic estimation still remains a potentially powerful and attractive tool. There have been many reported empirical evaluations both on actual and simulated data sets of synthetic estimation in recent years, including Levy (1971), Gonzalez (1973), Gonzalez and Hoza (1978), and Schaible (1979). Several of these types of studies are described in a volume edited by Platek and Singh (1986).

Royall (1970, 1973), using a model-based approach, also considered the problem of estimating totals in finite populations, when auxiliary information is available. He established a probability model of the relationship between the variable of interest and the auxiliary variable and then derived optimal subdomain predictors.

Holt, Smith and Tomberlin (1979) and Laake (1979) applied the predictive approach of Royall to the problem of small area estimation. Laake (1979) found that in contrast to the synthetic approach, where biased estimators are usually obtained without an explicit method of estimating the bias, the prediction approach yielded estimates of mean squared error (MSE) as a tool for the comparison of estimators. In the problem of estimating small area totals, Holt, Smith and Tomberlin (1979) specified various possibilities of population structure in order to model the assumed relationship across subareas. With a specified model, it becomes possible to determine whether or not it is supported by the data and also to study the effect of model misspecification on the bias of the observed estimators. Under different models, the variance of the estimator, the estimate of the variance and MSE change. They built model-based confidence intervals, which have interpretations in terms of repeated realizations under the super-population model.

Purcell and Kish (1979, 1980) reviewed the different existing techniques of small area estimation, subdividing them into the following broad categories, regression-based procedures, the use of empirical Bayes and of Bayesian methods, superpopulation prediction theory, clustering techniques, and categorical data analysis methods. They underlined the fact that small area domain estimation should not be considered as a homogeneous problem, but that there exist many other interacting factors such as domain size which should be taken into account when choosing the type of estimator. Särndal (1984) later confirmed this.

The most serious shortcoming of model-dependent estimators is that useful estimates of mean squared errors are not available using fixed effects models because associated variance estimates do not reflect the bias inherent in estimates based on models having a reduced set of parameters. Two different approaches were then taken to the problem of small area estimation.

Fay and Herriot (1979) used the James-Stein theory of estimation (James and Stein 1961) on sample data to determine estimates of income for small places from the 1970 US Census of Population and Housing. In fact, they used an empirical Bayes approach which originated with Robbins (1955) and has been described by Efron and Morris (1975), thus formalizing the meritorious suggestion of Madow and Hansen (1975) of forming a weighted average of the sample and regression estimates. A similar approach by Schaible, *et. al.* (1977) gives a method for arriving at a composite estimator which is the weighted average of the unbiased and synthetic estimators. For other examples of empirical Bayes methods for small area estimation based on standard normal theory see Stroud (1987) and Cressie (1988).

Battese, Harter and Fuller (1988), using a prediction approach, proposed a nested error regression model in order to estimate means. A more general model, a random coefficients regression model, had been previously proposed for a similar problem by Dempster, Rubin and Tsutakawa (1981). They used Bayesian techniques to estimate fixed and random effects in covariance component models when the covariances and variances are tentatively assumed to be known and the EM algorithm to subsequently estimate these unknown parameters. The introduction of random effects models not only allows for standard maximum likelihood estimation, but also provides measures of the reliability of the final estimates of the parameters in the form of posterior variances.

Ericksen (1980) suggested using the mean squared error (MSE) to evaluate effectiveness of regression in small area estimation. He attempted to answer such questions as: When should more predictor variables be added to the regression equation? Should James-Stein weighting procedures be used when the synthetic and the regression estimate are far apart? He also warned of the effects of outliers on both the resulting estimate and its estimated error. Perhaps the effect on small area estimators of the failure of the linear model assumptions should be more seriously studied.

Although applied to the estimation of counts such as unemployment and mortality statistics, most of these techniques described were designed primarily for continuous outcome variables. Purcell and Kish (1980) introduced a categorical data analysis method for obtaining estimates of counts for small domains. Essentially, their methodology involves fitting log-linear models to the data, omitting some of the higher order interaction terms and obtaining estimates by the iterative proportional fitting algorithm described by Deming and Stephan (1940). We propose to extend these models to the problem of estimation of proportions in small domains as originally conceived by Dempster and Tomberlin (1980) by applying empirical Bayes techniques to logistic regression models with random effects. This would have the added advantage that a measure of uncertainty of the small area estimates would be available through the approximate posterior variances. The estimator proposed here is similar in nature to the composite one used by Schaible *et. al.* (1977) for unemployment rates, the principal difference being in the method for choosing the weights. We feel, however, that the empirical Bayesian paradigm gives a more natural and intuitive method for determining the weights. Empirical Bayes estimation based on simple logistic random effects has already proven useful in studying regional variation in mortality rates by Miao (1977). Somewhat more complex random effects models have been used for proportions on data from the World Fertility Survey (Wong and Mason, 1985) and for Poisson parameters on automobile insurance data (Weisberg, Tomberlin, and Chatterjee 1984 and Tomberlin 1988).

Roberts, Rao and Kumar (1987) fitted logistic regression models to binary outcome data obtained using complex sampling schemes, constructed "pseudo-maximum likelihood" estimators, and compared their estimates to unbiased ones. They also proposed a goodness-of-fit test for their model, which takes the sampling design into account. A fundamental difference between our approach and that of Roberts, *et. al.*, is that by incorporating the characteristics of the sample design into the model, we can estimate parameters, and obtain readily interpretable measures of their reliability by means of standard maximum likelihood techniques.

2. THE MODEL

Following the framework of Dempster and Tomberlin (1980), in its most general form, we specify a model which describes the probabilities associated with individuals in the population as a function of categorical variables, continuous covariates and sampling characteristics. The models we consider in this paper are specific examples of the following,

$$\text{logit} (\pi_{\mu\nu}) = \theta_{\mu} + \mathbf{X}_{\mu\nu} \underline{\beta} + \phi_{\nu} \quad (2.1)$$

where $\pi_{\mu\nu}$ represents the probability of a "response" for the ν -th unit in the μ -th cell, the subscript μ refers to a set of categorical variable covariates, and the subscript ν refers to a set of nested sampling characteristics, indicating PSU, SSU within PSU, and so on. The parameter θ_{μ} represents a sum of fixed classification effects, the parameter ϕ_{ν} represents a sum of random effects associated with sampling characteristics, the vector $\mathbf{X}_{\mu\nu}$ represents a vector of quantitative covariates, and the parameter $\underline{\beta}$ is a vector of fixed logistic linear regression parameters. The random effects parameters are assumed to have some parametric distribution, usually a multivariate normal distribution. The probabilities $\pi_{\mu\nu}$ are obtained by inverting the logit transformation as follows,

$$\pi_{\mu\nu} = [1 + \exp\{- (\theta_{\mu} + X_{\mu\nu} + \phi_{\nu})\}]^{-1}. \tag{2.2}$$

For purposes of illustration, consider the following simple example. Let the proportion of interest be the labour force participation rate. Suppose we have one classification variable indicating sex and one continuous covariate indicating the age of the individual. Suppose further that the sample design is a simple, two stage cluster sample. In the first stage, a sample of counties is drawn and simple random samples of individuals within selected counties are drawn at the second stage.

For estimation purposes, consider the following model,

$$\text{logit} (\pi_{\mu\nu}) = \theta_{\mu} + X_{\mu\nu} \beta + \phi_i \tag{2.3}$$

$$\phi_i \sim \text{i.i.d. Normal} (0, \sigma^2). \tag{2.4}$$

Here, the classification subscript, μ , indicates the sex of the individual; the sampling characteristics subscript, $\nu = ij$, indicates the j -th individual within the i -th PSU; $X_{\mu\nu}$ indicates the age of the individual and ϕ_i is a random effect associated with the i -th PSU.

The consequence of assuming that the PSU effects are independent, identically distributed is that PSU departures away from the fixed part of the model are treated as exchangeable; that is, apart from effects of age and sex, no systematic information exists regarding differential employment rates among the counties in the population. Obviously in a realistic situation, such information would exist, for example, dominant industry, distance from principal markets, retail sales, etc. In such cases, this auxiliary information should be incorporated into the model. However, for purposes of illustration, we will continue with this simple model. The choice of a normal distribution of the error terms is a mathematical convenience, and the consequences of this choice must also be evaluated after actual data analysis. Extensions from the simple model described in (2.3-4) to include additional covariates, both categorical and quantitative is straight forward.

In theory, extensions to the model allowing for more complex sample designs is also simple. For example, data drawn using a three stage sample could be modelled using nested random effects as follows.

$$\text{logit} (\pi_{\mu\nu}) = \theta_{\mu} + X_{\mu\nu} \beta + \phi_i + \phi_{j(i)} \tag{2.5}$$

$$\phi_i \sim \text{Normal} (0, \sigma_1^2)$$

$$\phi_{j(i)} \sim \text{Normal} (0, \sigma_2^2).$$

Here, the sampling characteristics subscript, $\nu = ijk$ refers to the k -th individual within the j -th SSU within the i -th PSU. The parameter ϕ_i is the random effect associated with the i -th PSU, and $\phi_{j(i)}$ is the nested random effect associated with the j -th SSU within the i -th PSU. Stratification variables could also be incorporated within the fixed effects part of the model. While it is simple to write down the models corresponding to sample designs with several stages, without further research, it is not yet clear how difficult it will be to produce estimates based on these more complex models.

In an actual application, it would be necessary to use the data to identify predictor variables. This would require the development of some sort of model selection techniques. While not the primary focus of this paper, one might conceive of such a technique being based on an initial analysis using conventional variable selection techniques for logistic regression models as described by Haberman (1978), for example. Such an analysis could be conducted, ignoring the random effects parameters. Having chosen a set of predictors, the random effects would then be incorporated in the manner dictated by the sample design.

3. ESTIMATES

In this section, we develop empirical Bayes estimates for the simple model described in equations (2.3-4). First, it is assumed that the variance component, σ^2 , is known, and Bayesian estimates of the probabilities $\pi_{\mu ij}$ are obtained. Then, the EM algorithm, as described by Dempster, Laird and Rubin (1977), is used to obtain the maximum likelihood estimate of σ^2 allowing for empirical Bayes estimates. Finally, posterior variances of these estimates are obtained. The development of these estimates is similar to that described by Laird (1978) and by Tomberlin (1988).

3.1 Bayes Estimates

As noted by Laird (1978) in her analysis of contingency tables, by Dempster, Rubin and Tsutakawa (1981) in their analysis of variance components for linear models, and by Tomberlin (1988) in his analysis of Poisson data, a Bayesian analysis of a mixed model such as described in (2.3-4) can be obtained by placing a flat prior on the fixed parameters, θ_μ and β and the proper prior given in (2.4) on the random parameters, ϕ_i .

Let the vector of 0-1 outcome variables indicating membership in the labour force be represented by y and let π represent a vector of the individual probabilities $\pi_{\mu ij}$. The data are then distributed as a product binomial given by,

$$p(y | \pi) \propto \prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})}. \quad (3.1)$$

The prior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta | \sigma^2) \propto \exp \left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right]. \quad (3.2)$$

Thus, the joint distribution of the data, y , and the parameters is given by,

$$\begin{aligned} p(y, \underline{\theta}, \underline{\phi}, \beta | \sigma^2, \mathbf{X}) &= p(y | \underline{\theta}, \underline{\phi}, \beta, \sigma^2, \mathbf{X}) p(\underline{\theta}, \underline{\phi}, \beta | \sigma^2, \mathbf{X}) \\ &\propto \left[\prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})} \right] \exp \left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right]. \end{aligned} \quad (3.3)$$

From this, the posterior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta \mid \mathbf{y}, \sigma^2, \mathbf{X}) \frac{p(\mathbf{y}, \underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathbf{X})}{p(\mathbf{y} \mid \sigma^2, \mathbf{X})}. \tag{3.4}$$

It is not feasible to obtain a closed form expression for the posterior given in (3.4) due to the intractable integration required to obtain the marginal distribution of \mathbf{y} . Here we adopt the approximation employed by Laird (1978) and by Tomberlin (1988). The posterior is expressed as a multivariate normal distribution having its mean at the mode of (3.4) and covariance matrix equal to the inverse of the information matrix evaluated at the mode.

Obtaining the mode requires solving the following set of equations. This can be accomplished by using a multivariate Newton-Raphson algorithm.

$$\sum_{\mu ij} y_{\mu ij} X_{\mu ij} = \sum_{\mu ij} \hat{\pi}_{\mu ij} X_{\mu ij} \tag{3.5}$$

$$\sum_{ij} y_{\mu ij} = \sum_{ij} \hat{\pi}_{\mu ii} \tag{3.6}$$

$$\sum_{\mu j} (y_{\mu ij} - \hat{\pi}_{\mu ij}) - \frac{\hat{\phi}_i}{\sigma^2} = 0. \tag{3.7}$$

The posterior covariance matrix of the parameters is found by inverting the negative of the second derivative matrix of the log of (3.4) taken with respect to the parameters, and evaluated at the mode. Note that neither the equations for the mode, nor the covariance matrix involve the intractable denominator of (3.4).

Elements of the inverse of the posterior covariance matrix are given by,

$$\frac{-\partial^2}{\partial \beta^2} = \sum_{\mu ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) X_{\mu ij}^2 \tag{3.8}$$

$$\frac{-\partial^2}{\partial \theta_{\mu}^2} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \tag{3.9}$$

$$\frac{-\partial^2}{\partial \phi_i^2} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) - \frac{1}{\sigma^2} \tag{3.10}$$

$$\frac{-\partial^2}{\partial \beta \partial \theta_{\mu}} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) X_{\mu ij} \tag{3.11}$$

$$\frac{-\partial^2}{\partial \beta \partial \phi_i} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) X_{\mu ij} \tag{3.12}$$

$$\frac{-\partial^2}{\partial \theta_{\mu} \partial \phi_i} = \sum_j \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}). \tag{3.13}$$

3.2 Empirical Bayes Estimates

To obtain empirical Bayes estimates, the prior variance, σ^2 , must be estimated from the data. A reliable estimate requires a reasonable number of PSU's in the sample; otherwise, if the number of PSU's is too small, a purely Bayesian approach is recommended. We propose to estimate the prior variance using an EM algorithm as described by Dempster, Laird and Rubin (1977). The general framework for the estimates is similar to that employed by Laird (1978) for contingency table analysis, and Tomberlin (1988) for Poisson data in a two way classification. The estimates for the simple two-stage sample are obtained in exactly the same way as used by Leonard (1988).

The algorithm is initiated by choosing a starting value, $\sigma_{(0)}^2$, for the variance component. The posterior distribution of the random effects, ϕ_i , is then obtained by carrying out a Bayesian analysis as described in Section 2. This posterior distribution is then used to implement the E-step. The expected value of the sufficient statistic is calculated conditional on the data. The M-step is then completed by merely calculating the maximum likelihood function of the sufficient statistics. For a more complete description of the EM algorithm for regular exponential densities, see Dempster, Laird and Rubin (1977). The process is then repeated with a Bayesian analysis based on the updated estimate of the variance component, $\sigma_{(1)}^2$. The algorithm is continued until it converges.

3.3 Estimates of Small Area Proportions

Estimates together with associated posterior variances and covariances for parameters of the model given in (2.3-4) are presented in Sections 3.1 and 3.2. These estimated parameters are then employed to obtain estimates for small area proportions using a predictive approach. Assuming that the sample sizes within each area are small compared to those of the corresponding populations, this can be accomplished by averaging the individual estimated probabilities:

$$\hat{p}_i = \frac{\sum_{\mu j} \hat{\pi}_{\mu ij}}{N_i} \quad (3.14)$$

where N_i is the number of individuals in the i -th small area, and where the estimated probability associated with the μij -th individual, $\hat{\pi}_{\mu ij}$ is obtained by inverting the logistic function as follows,

$$\hat{\pi}_{\mu ij} = [1 + \exp\{-(\hat{\theta}_\mu + X_{\mu ij}\hat{\beta} + \hat{\phi}_i)\}]^{-1}. \quad (3.15)$$

To develop posterior variances for the estimates of small area proportions, it is convenient to adopt a more conventional notation for the linear part of the model, using dummy variables to indicate classifications. Let $Z_{\mu ij}$ represent a vector of predictor variables, both quantitative and qualitative, associated with the μij -th individual and let $\underline{\Gamma}$ represent a vector of the parameters of the model. Then,

$$Z_{\mu ij}^T \underline{\Gamma} = \theta_\mu + X_{\mu ij} \beta + \phi_i, \quad (3.16)$$

$$\hat{\pi}_{\mu ij} = [1 + \exp (-Z_{\mu ij}^T \hat{\underline{\Gamma}})]^{-1}. \tag{3.17}$$

Then, using a standard Taylor Series method, the posterior variance of the estimated small area proportion can be approximated as,

$$\text{Var}(\hat{p}_i) = \left[\sum_{\mu j} Z_{\mu ij}^T \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \right] \frac{\hat{\underline{\Sigma}}_{\Gamma}}{N_i^2} \left[\sum_{\mu j} Z_{\mu ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \right]. \tag{3.18}$$

Here, $\hat{\underline{\Sigma}}_{\Gamma}$ is the posterior covariance matrix of the estimated logistic regression parameters $\hat{\underline{\Gamma}}$.

Should the samples within small areas be substantial parts of the associated populations within those areas, then some additional gains in precision could be made by predicting only for the non-sampled units, in the spirit of the finite population sampling prediction methods originally described by Royall (1970).

4. THE SIMULATION STUDY

A simulation study was carried out to illustrate the characteristics of three different methodologies for producing local area estimates of proportions. The three methods evaluated were, the classical unbiased estimates, model-based estimates similar to the straightforward “synthetic estimates” of Gonzalez and Hoza (1978), and a modification of the proposed empirical Bayes estimates described in section 3, above. Data were simulated for a two-stage sample design. The 15 primary sampling units (PSU’s) were also treated as the local areas for which individual estimates of labour force participation rates were required. Within each of the 15 PSU’s, simple random samples of 25 individuals were drawn, for a total sample size of 375. The local area populations were assumed to be infinite so that complications associated with finite population sampling could be avoided.

As evaluations for local area estimates were required, it was decided to simulate resampling at the second stage only. That is, the same 15 PSU’s were drawn for each of the simulation studies. Each replicate consisted of a different sample drawn within these PSU’s. The study was based on 205 replications.

The data were generated using the model described in equation (2.3). The parameters were defined as follows,

$$\begin{aligned} \theta_1 &= -0.5 \\ \theta_2 &= -1.0 \\ \beta &= 0.1. \end{aligned} \tag{4.1}$$

The random parameters ϕ_i were generated from a normal distribution having mean zero and standard deviation 0.25. The $\pi_{\mu\nu}$ were obtained by inverting the logistic transformation as given in equation (3.15).

Here, θ_1 and θ_2 are the fixed effects associated with men and women respectively. That is, the odds ratio for labour force participation of men to that of women is $\exp[0.5] = 1.65$. The parameter β is the slope parameter associated with age, and the ϕ_i are the logistic random effects associated with the 15 PSU’s, or local areas.

Table 1
Population Labour Force Participation Rates by Local Area

Local Area	1	2	3	4	5	6	7	8
Participation Rate	0.79	0.79	0.96	0.88	0.90	0.95	0.86	0.96
Local Area	9	10	11	12	13	14	15	
Participation Rate	0.61	0.87	0.81	0.91	0.94	0.92	0.83	

The predictor variables, were generated with identical distributions for each of the 15 local areas. Age was distributed uniformly on the interval 20 to 40 years, the sex of each individual was drawn from a Bernoulli distribution with proportion 0.5, and the two predictor variables were assumed to be independently distributed. The population labour force participation rates for the 15 local areas are displayed in Table 1. As each local area was assumed to have the same distribution on the predictor variables, the only source of variation from area to area was the random local area effects, the ϕ_i . The random nature of these effects can produce a substantial variation in local area participation rates as is particularly evidenced by local area 9.

The observed local area sample proportions were used as unbiased estimates. The synthetic estimator was based on the following fixed effects, logit model,

$$\text{logit}(\pi_{\mu\nu}) = \theta_{\mu} \tag{4.2}$$

where, $\pi_{\mu\nu}$ and θ_{μ} are defined as for the random effects model in (2.3). Notice, only data from a particular local area are used to form the unbiased estimator while data are pooled from all local areas to obtain the synthetic estimator. However, the synthetic estimators will be biased to a degree which depends on the extent that model (4.2) fails to capture differences between local areas.

The third estimator studied here is a modification of the proposed empirical Bayes estimator described in Section 3. Due to the amount of computer time required to estimate the variance component associated with the local area effects, in fact, the Bayes estimator described in Section 3.1 was employed. The prior variance used for these estimates was the known value of the variance given in (4.1) used to simulate the data. As a result of this compromise, the results for the “empirical Bayes” estimator given below would be expected to be somewhat better than those which would be obtained using a true empirical Bayes estimator. However, sensitivity analyses aimed at determining the effect of changes in the prior variance indicate that the results which would be obtained using the empirical Bayes estimator would not be expected to substantially differ from those reported here for the modified estimator.

To look at bias, (in the classical sense of design-based inference) the estimates were averaged over all 205 replicates. Averages for each of the 15 local areas, for each estimation method are presented in Figure 1. The population rates are plotted as the “True Proportions”. These rates are almost exactly the same as the average unbiased estimates, and for the most part, are not visible on the graph. This confirms the unbiased nature of the classical estimates.

The synthetic estimates do not vary much from local area to local area. As each local area rate is based on the same pooled, fixed parameter estimates, the only source of variability from

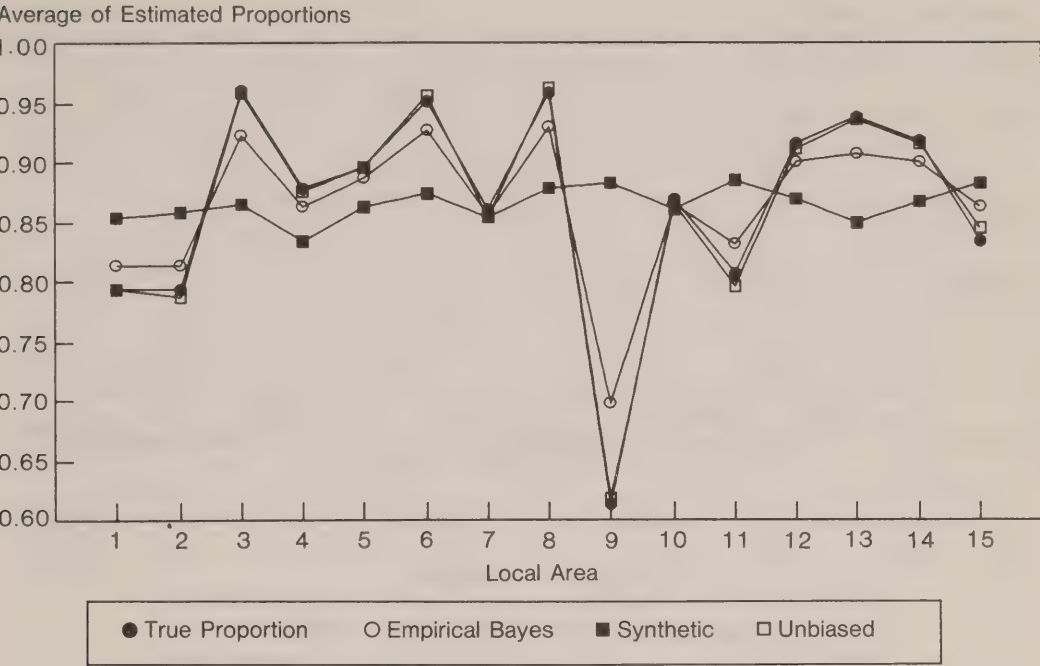


Figure 1. Averages of the estimated labour force participation rates for each of the three estimation methods plotted by local area

local area to local area is the small variability in the realized distributions of the predictor variables. The bias of this estimator can be large, as for example is the case for local area 9, where the synthetic method has a large positive bias. On the other hand, it should be noted that the synthetic method could not be expected to perform very well where there is little variability between the local area distributions of predictor variables.

The averages of the proposed estimates are in between the two extremes of the unbiased and synthetic estimates. They are biased, again in the classical sense, but their biases are smaller than those of the fixed effects model synthetic estimators.

Empirical Root Mean Square Errors (RMSE) were also calculated for each of the three estimators. These are presented in Figure 2. This plot demonstrates graphically where the synthetic estimator performs well and where it performs poorly. For local areas 7 and 10, where the local area effect is close to zero, the expected value of the synthetic estimator is very close to the population proportion. In these areas, the synthetic estimator has by far the smallest RMSE. By pooling data from the whole sample, it obtains a small sampling variance. On the other hand, in local area 9 where the local area effect is quite large, the associated RMSE for the synthetic estimator is also very large, due to its large bias. The modified empirical Bayes estimator obtains most of the reduction in RMSE that results from pooling the data across local areas, without suffering from the large bias associated with the synthetic estimator in those areas with large local area effects. In all but two cases, the modified empirical Bayes estimator achieves a smaller RMSE than the unbiased estimator. For local area 3, the RMSE's for the two estimators are about the same, and for local area 9, with a large local area effect, that of the modified empirical Bayes estimator is somewhat larger than that of the unbiased estimator. In short, the modified empirical Bayes estimator is sometimes the best of the three and never the worst.

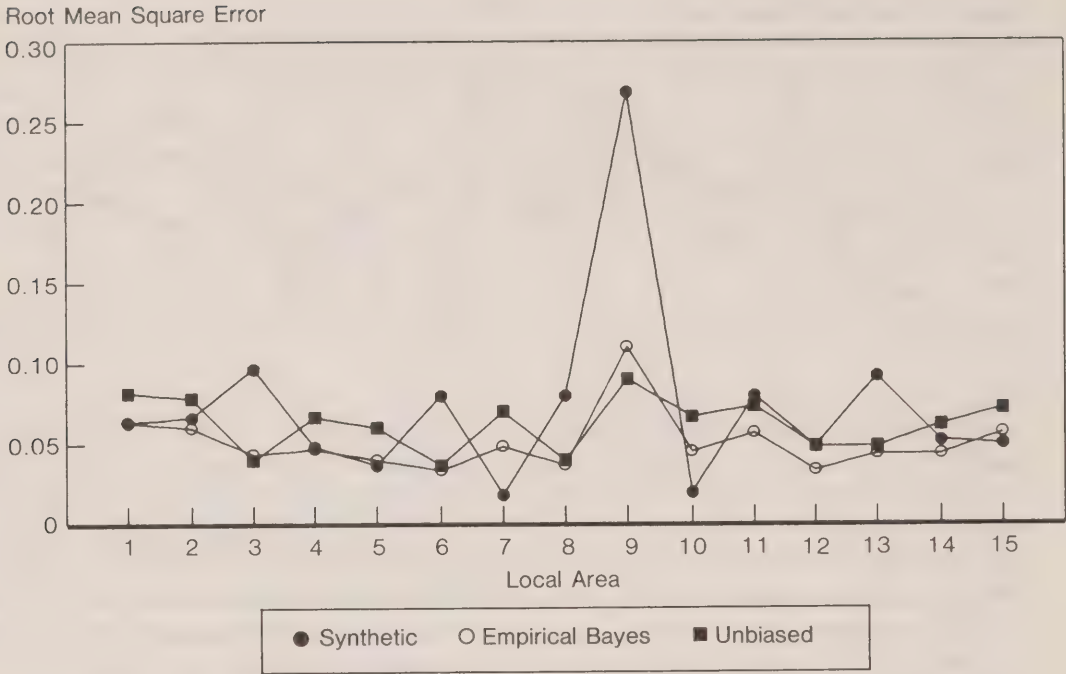


Figure 2. Empirical Root Mean Square Errors associated with each of the three estimation techniques plotted by local area

One of the principal shortcomings of the usual, fixed effects synthetic estimators is the difficulty in obtaining useful measures of associated accuracy. One can only obtain measures of sampling variances. Measures of bias which reflect model inadequacies are not available. For unbiased estimates, on the other hand, the usual estimates of sampling variability are also mean square error estimates as there is no bias. For empirical Bayes estimates, measures of uncertainty are available from the posterior covariance matrix of the parameters. These posterior variances reflect sampling variability as well as the “bias” which comes from simple fixed effects model inadequacies. This latter source of uncertainty is captured via the variability in the local area effects parameters.

The usefulness of these measures of uncertainty are compared graphically in Figure 3. The vertical axis corresponds to the empirical root mean square error (RMSE) which is obtained by comparing the individual replicate estimates with the known population proportions for each local area. The horizontal axis corresponds to the “reported RMSE”. For the classical unbiased estimates, these are merely the sampling standard deviations for simple random sampling. For the synthetic estimates, they are also sampling standard deviations, corrected for the cluster sampling. The “reported RMSE” for empirical Bayes estimates are the square roots of the posterior variances of the estimated proportions which were obtained using the methods described in Section 3.2 above.

Note that the points corresponding to the unbiased estimates lie along a line indicating that the reported RMSE’s are very close to the empirical RMSE’s. This is as expected since there is no bias in these, so the reported RMSE’s and the empirical RMSE’s are merely sampling standard deviations. As opposed to this, the points corresponding to the synthetic estimates are in a cluster above 0.015 to 0.020 on the horizontal axes. For these estimates,

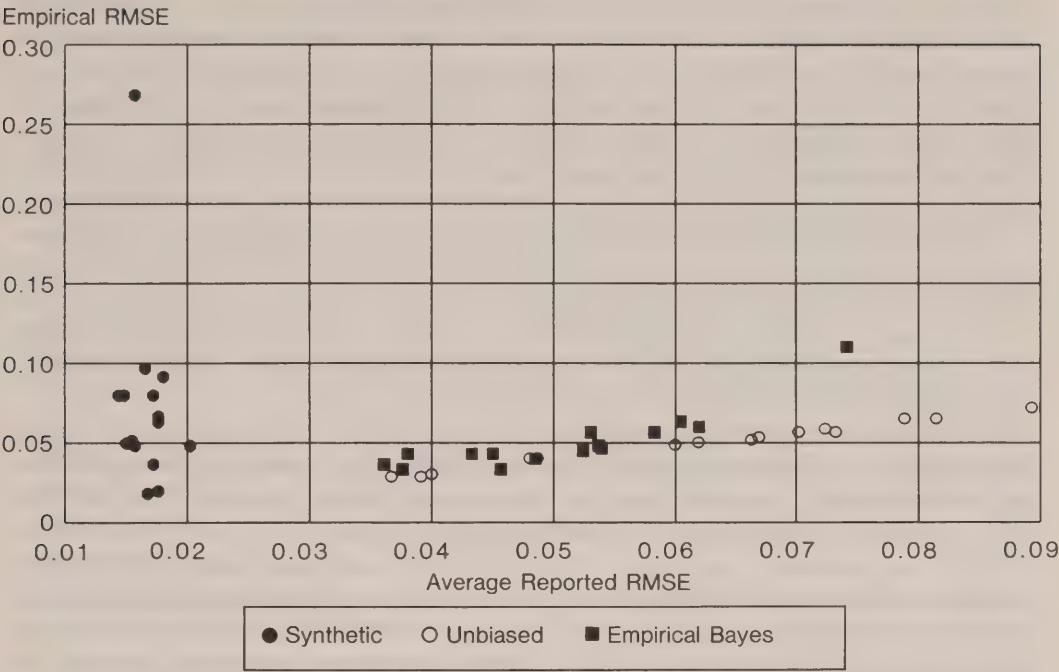


Figure 3. Empirical Mean Square Error vs “Reported Mean Square Errors” for each of the three estimation techniques

the “Reported RMSE’s” are estimates of sampling standard deviations, which for these pooled estimates are all quite small. However, the empirical RMSE’s for these estimates are quite a different story. They range from 0.015 to 0.100, with one outlier in excess of 0.250 (local area 9). Sampling variances alone are not sufficient to describe the uncertainty associated with the estimates.

The case for the modified empirical Bayes estimators is again in between these two extremes. However, with respect to the relationship between reported RMSE and empirical RMSE it is much closer to the corresponding relationship for the unbiased estimators. With the exception of the point associated with local area 9, the average reported RMSE’s are very close to the corresponding empirical RMSE’s.

5. CONCLUSIONS

In the simple simulation of a two-stage sample where PSU’s correspond to local areas, the modified empirical Bayes estimators have been shown to be superior, overall to two standard alternatives. These have been evaluated in three ways, design-bias, root mean square error, and validity of estimable measures of uncertainty. The classical estimator is shown to be superior in terms of design-bias, as expected since it is design unbiased. In addition, valid estimates of RMSE’s are available using standard techniques. However, these estimators suffer from large RMSE’s due to the fact that they are formed from limited amounts of data. Indeed, unlike the other two alternatives, no estimates can be formed at all for local areas not in the sample.

At the other extreme, the synthetic estimator is far more stable than either of its competitors. Since all estimates are based on data from the whole sample, associated sampling variances are much smaller than those of the other two estimators. On the other hand, this estimator is unable to adjust for local areas which are quite different from the rest. This is the case, even when data are available in the sample that would indicate such a difference. As important, estimates of uncertainty in the form of sampling standard deviations for this estimator are particularly misleading since they are unable to account for departures from the fixed effects model.

As a compromise between these two estimators, the modified empirical Bayes estimator performs well on all three assessments. By using the data from the specific local areas to the extent it is reliable, this estimator avoids the large biases associated with the synthetic estimator. On the other hand, by pooling information from the whole sample, it has smaller sampling variances than the unbiased estimator, and generally smaller RMSE's. Finally, posterior variances are available as useful measures of uncertainty.

Several tasks remain in the investigation of the proposed estimators. First, the effect of using true empirical Bayes estimators instead of modified ones must be assessed. Some guidelines for minimum number of sampling units for valid empirical Bayes inference are required. True empirical Bayes estimates employ estimated prior variances and methods which account for this additional uncertainty are required. For example, the bootstrap techniques investigated by Laird and Louis (1987) could be used. Second, the estimation techniques need to be generalized to handle three and more stages of sampling. While the theoretical extension is trivial, the computational implications are not. Finally, these techniques must be applied to real data before recommending their adoption as a standard alternative for local area estimation.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful suggestions of an associate editor and a referee, and the financial support of NSERC of Canada, and Concordia University.

REFERENCES

- BATTESE, G. E., HARTER, R. M., and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36.
- CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 191-208.
- DEMING, W. E., and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of The Royal Statistical Society, Ser. B*, 39, 1-38.
- DEMPSTER, A. P., RUBIN, D. B., and TSUTAKAWA, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.
- DEMPSTER, A. P., and TOMBERLIN, T. J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, 88-94.
- ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography* 10, 137-159.
- ERICKSEN, E. P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

- ERICKSEN, E. P. (1980). Can regression be used to estimate local undercount adjustments? *Proceedings of the Conference on Census Undercount*, 55-61.
- EFRON, B., and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.
- FAY, R. E., and HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 33-36.
- GONZALEZ, M. E., and HOZA, C. (1976). Small area estimation of unemployment. *Proceedings of the Section on Social Statistics, American Statistical Association*, 437-443.
- GONZALEZ, M. E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.
- GONZALEZ, M. E., and WAKSBERG, J. L. (1975). Estimation of the error of synthetic estimates. Unpublished paper presented at the first meeting of the International Association of Survey Statisticians, Vienna.
- HABERMAN, S. J. (1978). *Analysis of Qualitative Data Volume 1: Introductory Topics*. New York: Academic Press.
- HOLT, D., SMITH, T. M. F., and TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.
- JAMES, W., and STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 361-379.
- LAAKE, P. (1979). A predictive approach to subdomain estimation in finite populations. *Journal of the American Statistical Association*, 74, 355-358.
- LAIRD, N. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.
- LAIRD, N., and LOUIS, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association*, 82, 739-757.
- LEONARD, K. J. (1988). Credit scoring via linear logistic models with random parameters. Ph. D. Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montréal.
- LEVY, P. S., (1971). The use of mortality data in evaluating synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 323-331.
- LEVY, P. S., and FRENCH, D. K. (1978). Estimation of health characteristics. *Vital and Health Statistics*, Ser. 2, No. 75, NCHS, Washington, DC.
- MADOW, W. G., and HANSEN, M. H. (1975) On statistical models and estimation in sample surveys. Contributed Papers, 40th Session of the International Statistical Institute, Warsaw, Poland, 554-557.
- MIAO, L. L. (1977). An empirical Bayes approach to analysis of inter-area variation, Ph. D. Dissertation, Department of Statistics, Harvard University.
- MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-54.
- O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.
- PLATEK, R., and SINGH, M. P. (1986). *Small Area Statistics--An International Symposium '85* (Contributed Papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University-University of Ottawa, Canada.
- PURCELL, N. J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

- PURCELL, N. J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.
- ROBERTS, G., RAO, J. N. K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROBBINS, H. I. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium*. Berkeley: University of California Press, 157-164.
- ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALL, R. M. (1973). Discussion of papers by Gonzalez and Ericksen. *Proceedings of the Section on Social Statistics, American Statistical Association*, 42-43.
- SÄRNDAL, C. E. (1984). Design consistent versus model dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCHAIBLE, W. L. (1979). A composite estimator for small area statistics. In *Synthetic Estimates for Small Areas* (NIDA Research Monograph 24), edited by J. Steinberg. Rockville, MD: National Institute on Drug Abuse, 36-53.
- STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics*, (Eds. R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh.). New York: Wiley, 124-137.
- TOMBERLIN, T. J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.
- U.S. National Center for Public Health Statistics (1968). *Synthetic State Estimates of Disability*, PHS Publication No. 1759.
- WEISBERG, H. I., TOMBERLIN, T. J., and CHATTERJEE, S. (1984). Predicting insurance losses under a cross-classification: a comparison of alternative approaches. *Journal of Business and Economic Statistics*, 2, 170-178.
- WONG, G. Y., and MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

Updating Size Measures in a PPSWOR Design

ALAN SUNTER¹

ABSTRACT

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities may now be regarded as being proportional to the new size measures. The method described in this article differs from methods already described in the literature in that it is valid for any sample size and does not require enumeration of all possible samples. Further, it does not require that the old and the new sampling methods be the same and hence it provides a convenient way not only of updating size measures but also of switching to a new sampling method.

KEY WORDS: PPSWOR; Sample updating; PPS sequential sampling.

1. INTRODUCTION

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. This occurs, for example, when the psu's are census enumeration areas (or collections of census enumeration areas) and a new census has made new population/housing counts available or when, because of observed uneven growth in EA populations in an intercensal period, it is decided to do an interim update of size measures in a sampling stratum. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities, originally proportional to the old size measures, may now be regarded as being proportional to the new ones. A comprehensive treatment of the problem for $n = 1$ is given by Kish and Scott (1971) and is itself a generalization of a method given earlier by Keyfitz (1951). They point out that their method may be extended without difficulty to with replacement sampling (PPSWR) for $n > 1$. Their method may also be used (Drew, Choudhry, and Gray 1978; Platek and Singh 1978) for $n > 1$ when the PPSWOR procedure used is that due to Rao, Hartley and Cochran (1962), since this method involves the formation of n random groups and subsequent selection of a single psu from each group. It breaks down however if we wish, as indeed we probably would, to form new random groups according to the new size measures. Fellegi (1966) provides two methods applicable to a PPSWOR sample of $n = 2$ drawn by the Fellegi (1963) procedure.

The method given in this paper is similar to the second Fellegi method and, when applied to the examples in the Fellegi paper, gives very similar results. Unlike that method, however, it does not require the enumeration of all possible samples and hence is a feasible procedure for any value of n and N . Although it is formally applicable to any PPSWOR method for which it is feasible to calculate the selection probability of any sample selected it has its highest utility for PPSWOR methods in which all, or nearly all, n -tuple subsets are possible samples with

¹ Alan Sunter, President, A.B. Sunter Research Design & Analysis Inc., 63 Fifth Av., Ottawa, Canada, K1S 2M3.

probabilities approximately proportional to the product of their unit probabilities. The method of this type, used for purposes of illustration, is the author's pps sequential method (Sunter 1986, 1989).

2. REPLACEMENT PROCEDURE THEORY

We wish to reselect a PPSWOR sample, originally selected with probabilities $\{\pi_{11}, \pi_{12}, \dots, \pi_{1n}\}$ proportional to original size measures $\{z_{11}, z_{12}, \dots, z_{1n}\}$ under a new set of probabilities $\{\pi_{21}, \pi_{22}, \dots, \pi_{2n}\}$ proportional to new size measures $\{z_{21}, z_{22}, \dots, z_{2n}\}$. However, we want to do this in such a way that we have a high probability of retaining the original sample.

We assume that for any particular n -tuple S , including of course S' , the original sample actually selected, it is possible to calculate both $P_1(S)$, its selection probability under the original scheme, and $P_2(S)$, its selection probability under a new scheme. For many samples in many schemes (e.g. pps systematic) one or both of these probabilities may be zero although, obviously, $P_1(S')$ cannot be zero.

The procedure is as follows:

- Step 1: (a) Calculate $P_1(S')$, $P_2(S')$.
 (b) If $P_2(S') \geq P_1(S')$ then retain the sample.
 (c) If $P_2(S') < P_1(S')$ retain the sample with probability $P_2(S')/P_1(S')$. If rejected proceed to Step 2.
- Step 2: (a) If the original sample was not retained then draw a new sample, S_1 say, with probability $P_2(S_1)$. If $P_2(S_1) < P_1(S_1)$ then reject the sample, otherwise retain with probability $1 - P_1(S_1)/P_2(S_1)$. If rejected proceed to Step 2(b).
 (b) If the Step 2(a) sample was not retained then draw a new sample, S_2 say, and proceed as for Step 2(a).
 (c), (d), ... Repeat the Step 2(a), 2(b), ... procedure until a sample is retained.

The sample eventually retained by this process has the required probability structure for both unit probabilities and unit pair joint probabilities. In other words, it may be regarded as having been drawn under the new scheme. In particular, since it has the same joint probability structure, it has the same sampling variance.

Let P^* denote the probability that the process does not terminate at Step 1, P^{**} the conditional probability that it does not terminate at Step 2(a) given that it did not terminate at Step 1. Obviously P^{**} is then also the conditional probability that the process does not terminate at any subsequent step given that it did not terminate at any step preceding that step. We now have

$$\begin{aligned}
 P^* &= \sum_{i: P_2(S_i) < P_1(S_i)} (1 - P_2(S_i)/P_1(S_i)) P_1(S_i) \\
 &= \sum_{i: P_2(S_i) < P_1(S_i)} (P_1(S_i) - P_2(S_i)) \quad (1)
 \end{aligned}$$

where i now indexes the n -tuple subsets of the N population units, and

$$\begin{aligned} P^{**} &= 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &= 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (P_2(S_i) - P_1(S_i)) \end{aligned} \tag{2}$$

while, since $\sum_i P_1(S_i) = \sum_i P_2(S_i) = 1$, it is easy to see that the summation terms on the right of (1) and (2) respectively must be equal and we have $P^* = 1 - P^{**}$.

Denoting ultimate selection probability by P' we now have, by design:

For $i:P_2(S_i) < P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i) (P_2(S_i)/P_1(S_i)) \\ &= P_2(S_i), \text{ as required.} \end{aligned}$$

For $i:P_2(S_i) \geq P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i) + P^*(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*P^{**}(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^2(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^3(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + \dots \\ &= P_1(S_i) + P^*(P_2(S_i) - P_1(S_i))/(1 - P^{**}) \\ &= P_2(S_i) \end{aligned}$$

as required.

Finally, we observe that the expected number of Step 2 “trials”, given that the original sample was not retained at Step 1, is given by the binomial waiting time distribution as $1/(1 - P^{**}) = 1/P^*$.

3. APPLICATION AND EXAMPLES

The new scheme need not be the same (even apart from the change in unit probabilities) as the old one. We could switch, for example, from a sample originally drawn under pps systematic sampling to one drawn under the author’s (Sunter 1986, 1989) pps sequential scheme or even from PPSWR (pps with replacement) to a PPSWOR scheme. In the latter case, of course, an original sample with multiple inclusions of a single psu has zero probability of selection in the new PPSWOR scheme. The procedure may still be used, it may be noted, even if we have included new psu’s in the stratum but are retaining the same sample size.

The procedure probably has its highest practical utility, as measured by its probability of retaining the same sample, when both the old and the new schemes are such that all, or nearly all, samples are possible and their probabilities are approximately proportional to the product of their unit selection probabilities. Under these circumstances, and provided that the changes in size measures are not extreme, $P_1(S_i)$ and $P_2(S_i)$ tend to have about the same values so that the probability of retaining the same sample will be relatively high. A practical PPSWOR method with the required properties is the author's, referred to above. Since we will use this method in the examples of the next section, we now describe it. There are two variants, in both of which we have to find a suitable ordering of the population and accumulate the size measures (which we assume to be scaled to sum to 1), in reverse order (so to speak), to give:

$$Z_i = \sum_{j=i}^N z_j; \quad i = 1, 2, \dots, N.$$

Variant 1: Order the population in any way such that

- (a) $nz_i \leq Z_i; \quad i = 1, 2, \dots, N - n$
- (b) $(n - i)z_i < Z_i; \quad i = n, n + 1, \dots, N - 1.$

Then select units until exactly n have been selected according to:

$$P(U_i | n_i) = \begin{cases} 1 & \text{if } n_i = N - i + 1 \\ n_i z_i / Z_i & \text{otherwise} \end{cases}$$

where n_i is the number of sample units still required to be selected when we arrive at the i -th population unit.

It is always possible to satisfy the ordering requirements (a) and (b). For example ordering by increasing size obviously satisfies both as does ordering by decreasing size down to the point (if any) at which (b) fails and then by increasing size. The latter ordering has some advantage in that it tends to minimize the slight (and, for practical purposes, negligible) deviation from strict pps for the last n units (see Sunter 1986). Variant 2 avoids these deviations altogether by taking advantage of the fact that if it occurs that there are $n_i + 1$ units remaining in the population for any i , then it is usually possible to simply discard one of these units with appropriate probability and retain the others.

Variant 2: Order the population in any way such that

- (a) $nz_i \leq Z_i; \quad i = 1, 2, \dots, N - n - 1$
- (b) $(n - i)z_j < Z_i; \quad j \geq i \geq N - n.$

Then

- (i) select according to $P(U_i | n_i) = nz_i / Z_i$ until either $n_i = 0$ or $n_i = N - i$, then
- (ii) if $n_i > 0$ discard one of the remaining units, say that indexed j , with probability $1 - n_i z_j / Z_i$ and select the others.

An algorithm for finding an ordering satisfying the requirements for Variant 2 is given in Sunter(1986) and is incorporated in the program used for the simulations of the next section. In both variants π_{ij} maybe calculated according to

$$\pi_{ij} = n(n - 1)z_iz_j\tau_{ij}$$

where $i < j$ (in the indexing of the ordering actually used) and

$$\begin{aligned}\tau_1 &= 1/Z_2 \\ \tau_i &= (1/Z_i + 1)(1 - z_1/Z_2) \dots (1 - z_{i-1}/Z_i).\end{aligned}$$

These expressions are exact for $i < j \leq N - n + 1$, and provide a very close approximation otherwise. They are easily calculated and give the method the advantage, unique among practical procedures for PPSWOR with $n > 2$, of the availability of variance estimation with negligible bias.

Pascal-like pseudocode for a routine that selects a sample according to Variant 1, at the same time calculating its probability and the value of τ_i for each selected unit, is given in an Appendix. It is easily extended to Variant 2 or modified to the calculation of $P(S)$ for an already selected sample.

3.1 Example 1

To illustrate these procedures we take first an example with $n = 2$ and $N = 4$, small enough for sample enumeration and manual calculation, where it will be seen that, in order to obtain the “new” size measures, we simply inverted the order of the original assignment. The Variant 2 ordering algorithm mentioned above gives (4,1,2,3) for the first set of size measures and (1,4,3,2) for the second. There are six possible samples, listed in column (1) of Table 2, whose probabilities under the Variant 2 algorithm are easily calculated, with results shown in columns (2) and (3). Column (4) gives the probability of retaining this sample at Step 1, given that it was the original selection. Column (5) gives the conditional probability of retention at any subsequent Step 2, given that no sample was retained at a preceding step.

It may be verified that the overall probability of retention of the same sample, given by the sum of the products of the values in columns (2) and (4), is 0.5465. This value may be compared with the overall probability of retention of the same sample when the new sample is selected independently, given by $\sum_i P_1(S_i)P_2(S_i) = 0.1168$. Thus even in this rather extreme example, we have considerably increased the likelihood of retaining the same sample.

Table 1
Selection Probabilities

PSU	z_{1i}	z_{2i}
1	0.15	0.35
2	0.20	0.30
3	0.30	0.20
4	0.35	0.15

Table 2

(1) Sample	(2) $P_1(S)$	(3) $P_2(S)$	(4) $P_{2 1}(S)$	(5) $P_{2 2}(S)$
1,2	0.0231	0.3231	1.0	0.9286
1,3	0.1154	0.2154	1.0	0.4643
1,4	0.1615	0.1615	1.0	0
2,3	0.1615	0.1615	1.0	0
2,4	0.2154	0.1154	0.5357	0
3,4	0.3231	0.0231	0.0715	0

3.2 Example 2

In a more realistic set of examples we now take $n = 4$ for a population of 100 psu's with "original" size measures independently assigned from the uniform or rectangular distribution $R(1,3)$. "New" size measures are assigned in a number of ways, described below. For these examples it is no longer feasible to enumerate all possible samples or to perform the sample selection and sample probability calculations manually. However, writing a computer program to do the latter and to apply the reselection procedure was a straightforward task. The program was used to perform 200 iterations, for each example, of selection of a sample using Sunter's Variant 2 with probabilities proportional to the first set of size measures with subsequent application of the procedures described above for reselection of a sample with probabilities proportional to the second set of size measures. The program, running on an XT-compatible operating at 7.16 MHz, generated and sorted the populations of size measures and performed 200 iterations of the sample selection and reselection in about three minutes.

Case 1, in which we have assigned new size measures from the same distribution independently of their original values, may be seen as a "worst practical case" scenario. Case 2, in which 10% of the psu's have doubled in size with the rest remaining unchanged, is an approximation of a "scattered development" scenario. Case 3 illustrates the random perturbation of size measures by an amount rectangularly distributed over an interval equal to the original size measure. From Table 3 it may be seen that with probabilities ranging from 0.67 in the "worst case" scenario to 0.81 in the "scattered development" scenario, we retain the original sample. For those cases in which the original sample is rejected the average number of Step 2 trials required to select a new sample agreed closely with the predicted value of $1/P^*$.

Table 3

200 Iterations of a Size Measure Update Procedure, $n = 4, M = 100$;
Original Size Measures from $R(1,3)$

Case	Source of π_{2i}	Step 1 Retentions	Average Step 2 Trials	Estimated P^*
1	$z_{2i} \approx R(1,3)$	134	2.98	0.33
2	$z_{2i} = 2 \cdot z_{1i}$ for 10% of psu's	153	5.53	0.19
3	$z_{2i} = R(z_{1i}/2, 3z_{1i}/2)$	154	4.17	0.25

ACKNOWLEDGEMENTS

The author would like to thank the Editor and two referees for helpful comments and the correction of errors in the original presentation.

APPENDIX

Pseudocode for Variant 1 of PPS Sequential Sampling

It is assumed here that the population of size measures has already been given a suitable ordering, say by the algorithm given in Sunter (1986) and that its index, i , in this ordering identifies the unit. Size measures, scaled to sum to 1, are stored in an array $z[1 \dots \text{PopSize}]$ with their cumulative values (accumulated from PopSize down to 1) stored in an array $Z[1 \dots \text{PopSize}]$. The meaning of the variables will be clear from the names that they are given. The results are to be stored in an array $\text{Sample}[1 \dots \text{SamSize}, 1 \dots 3]$ in which the elements are population index i , unit probability π_i , and τ_i respectively. "Random" is a function that returns a random number uniformly distributed on the interval (0,1). The indentations in the code written below are intended to facilitate the visual pairing of the begin/end's that delineate a compound statement.

{ Variables initialization }

$i = 1$; $\text{SamProb} = 1$; $\text{NumRem} = \text{SamSize}$; $\text{Gamma} = 1/Z[2]$;

{ Sampling routine }

while $\text{NumRem} > 0$ do

begin

if $i > 1$ and $i < \text{PopSize}$ then

$\text{Gamma} = \text{Gamma} * (1 - z[i-1]/Z[i]) * Z[i]/Z[i+1]$;

if $i = \text{PopSize} - \text{NumRem} + 1$ or $\text{Random} < = \text{Numrem} * z[i]/Z[i]$

then

begin

if $i < > \text{PopSize} - \text{NumRem} + 1$ then

$\text{SamProb} = \text{SamProb} * \text{NumRem} * z[i]/Z[i]$;

$\text{NumRem} = \text{NumRem} - 1$;

$\text{Sample}[\text{SamSize} - \text{NumRem}, 1] = i$;

$\text{Sample}[\text{SamSize} - \text{NumRem}, 2] = \text{SamSize} * z[i]$;

$\text{Sample}[\text{SamSize} - \text{NumRem}, 3] = \text{Gamma}$;

end else $\text{SamProb} = \text{SamProb} * (1 - \text{NumRem} * z[i]/Z[i])$;

$i = i + 1$;

end.

REFERENCES

- DREW, J.D., CHOUDHRY, G.H., and GRAY, G.B. (1978). Some methods for updating sample survey frames and their effects on estimation. *Survey Methodology*, 4, 225-263.
- FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- FELLEGI, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, 434-442.
- KEYFITZ, N. (1951). Sampling with probabilities proportional to size. *Journal of the American Statistical Association*, 58, 183-201.
- KISH, L., and SCOTT, A., (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, 24, 482-490.
- PLATEK, R., and SINGH, M.P. (1978). A strategy for updating continuous surveys. *Metrika*, 25, 1-7.
- SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.
- SUNTER, A.B. (1989). PPS Sampling in multistage designs: does it matter which method? Manuscript submitted to *Journal of Official Statistics*.

The Use of Administrative Records for Estimating Population in Canada¹

RAVI B.P. VERMA and RONALD RABY²

ABSTRACT

This paper examines the adequacy of estimates of emigrants from Canada and interprovincial migration data from the Family Allowance files and Revenue Canada tax files. The application of these data files in estimating total population for Canada, provinces and territories, was evaluated with reference to the 1986 Census counts. It was found that these two administrative files provided consistent and reasonably accurate series of data on emigration and interprovincial migration from 1981 to 1986. Consequently, the population estimates were fairly accurate. The estimate of emigrants derived from the Family Allowance file could be improved by using the ratio of adult to child emigrant rates computed from Employment and Immigration Canada's immigration file.

KEY WORDS: Interprovincial migration; Emigration; Population estimates; Census counts; Accuracy.

1. INTRODUCTION

The national Census, conducted every five years since 1951, provides a wide range of demographic data on the Canadian population. However, unlike some other industrialized countries, Canada does not have a continuous population registration to derive basic demographic data and track the movement of people over different geographic areas for non-census years. To fill this gap, since the 1940s Statistics Canada has developed a program of population and family estimates. For example, population estimates for Canada, provinces and territories, census divisions, and census metropolitan areas are produced using the latest census counts and several administrative data sources, including: Revenue Canada tax files and Family Allowance files for migration; Vital Statistics registration for births and deaths; and Immigrant Visa and Record of Landing Registration for immigration.

The strengths and weaknesses of these administrative files for estimating population and migration compared with 1981 Census data have been discussed elsewhere. (Statistics Canada 1987; Verma and Parent 1985; Norris, Britton and Verma 1982). In this paper, the accuracy of estimates of the components of population change for provinces and territories using the Family Allowance and Revenue Canada data sources will be evaluated by comparison with the 1986 Census counts. This evaluation will compare 1971, 1976 and 1981 data.

The paper is presented in the following sections: data sources and the methods of estimation; results of the evaluation; and conclusions and discussion.

2. DATA SOURCES AND THE METHODS OF ESTIMATION

This section describes the procedures for estimating total population, interprovincial migration, and emigration.

¹ Revised version of a paper presented at Statistics Canada Symposium on Statistical Uses of Administrative Data, November 1987.

² Ravi B.P. Verma and Ronald Raby, Demography Division, Statistics Canada, 4-A Jean Talon Building, Ottawa, Ontario, K1A 0T6.

2.1 Total Population

Quarterly and annual estimates of the total population of Canada and the provinces and territories, and annual totals for census divisions and census metropolitan areas, are produced by the component method. At the national level, the number of births and immigrants are added to, and the number of deaths and emigrants subtracted from, the base population (taken from the latest Census of Canada). By province and for smaller areas, estimates of internal migration are also taken into account.

The component method is expressed as follows:

$$\begin{aligned}\hat{P}(t + i) = P(t) + [B(t, t + i) - D(t, t + i) \\ + I(t, t + i) - E(t, t + i)] + N(t, t + i).\end{aligned}\quad (1)$$

Where, for any given province:

$\hat{P}(t + i)$ = estimate of population at time $t + i$

$P(t)$ = Census population counts at time t

B = number of births between time t and $t + i$

D = number of deaths between time t and $t + i$

I = number of immigrants between time t and $t + i$

E = number of emigrants between time t and $t + i$

N = number of net interprovincial immigrants between time t and $t + i$

$(t, t + i)$ = interval between the last census date and the reference date of the estimate.

2.2 Interprovincial Migration

Two administrative files are used to produce annual and quarterly estimates of interprovincial migration. Preliminary estimates are derived from Family Allowance files, while final figures are estimated from Revenue Canada income tax files.

2.2.1 Preliminary Estimates

The number of adult migrants is estimated using child migration figures derived from Family Allowance files, and ratios of adult out-migration rates to child out-migration rates ($f_{j,k}$) based on the most recent Revenue Canada tax file (calculated for 1 or 2 years before the reference date). Recipients of Family Allowance cheques must notify the Department of Health and Welfare of changes in address. These changes are compiled monthly for both province of origin and destination, by size of family (the number of children per family receiving the allowance). Coverage of the population by Family Allowance is comparable to that of the census (Statistics Canada 1987, p. 46). Estimates of the number of interprovincial out-migrants for all age groups are calculated as follows:

$$\hat{M}_{(j,k),18+} = \frac{M_{(j,k),0-17}}{P_{j,0-17}} \cdot f_{(j,k)} \cdot P_{j,18+} \quad (2)$$

$$f_{(j,k)} = \frac{M'_{(j,k),18+}}{\hat{P}_{j,18+}} \div \frac{M'_{(j,k),0-17}}{\hat{P}_{j,0-17}} \quad (3)$$

$$\hat{M}_{(j,k),0+} = \hat{M}_{(j,k),18+} + M_{(j,k),0-17} \quad (4)$$

where:

- $\hat{M}_{(j,k),0+}$ = estimated total number of persons out-migrating from province j to province k
- $\hat{M}_{(j,k),18+}$ = estimated number of adult out-migrants (aged 18+) from province j to province k
- $M'_{(j,k),18+}$ = number of adult out-migrants from province j to province k derived from Revenue Canada tax files
- $M'_{(j,k),0-17}$ = number of child out-migrants (aged 0-17) from province j to province k derived from Revenue Canada tax files
- $M_{(j,k),0-17}$ = number of child out-migrants from province j to province k , based on Family Allowance files
- $P_{j,18+}$ = estimated number of adults in province j , the difference between the total population estimates and estimates of the child population based on Family Allowance files
- $P_{j,0-17}$ = total number of children receiving Family Allowance payments in province j
- $f_{(j,k)}$ = estimation factor for adult migrants from province of origin j to province of destination k , based on estimates of migration from Revenue Canada tax files
- $\hat{P}_{j,18+}$ = number of adults in province j , Demography Division population estimates
- $\hat{P}_{j,0-17}$ = number of children in province j , Demography Division population estimates.

2.2.2 Final Estimates

Revenue Canada tax files are used to produce final estimates of interprovincial migrants. All individuals receiving an annual income above a specified minimum are required to file an income tax return by the end of April of each year. Migrant tax filers are identified by comparing area of residence from two consecutive tax returns. Information on the number and ages of dependents is imputed from the total amount of personal exemptions claimed by filers. An adjustment is made for segments of the population not covered by the Revenue Canada system; this includes people who neither file an income tax return nor appear as dependents in another filer's return (Norris and Standish 1983; Statistics Canada 1987).

2.3 Emigration

In Canada no system exists for recording emigrants; hence, their numbers must be estimated. Revenue Canada income tax files with an "out-of-Canada" address one year and an "in-Canada" address for the previous year are used to identify emigrants. The emigrant status of children under 17 years of age is determined from change of address notifications from Family Allowance recipients. By combining information from these two administrative files, both preliminary and final estimates of emigrants are generated. The estimation procedures are similar to those used to estimate preliminary interprovincial migration:

$$\hat{E}_j = \left[\frac{E_{j,0-17}}{P_{j,0-17}} \cdot f_c \cdot P_{j,18+} \right] + E_{j,0-17} \quad (5)$$

$$f_c = \frac{E'_{c,18+}}{\hat{P}_{c,18+}} \div \frac{E'_{c,0-17}}{\hat{P}_{c,0-17}} \quad (6)$$

$$\hat{E}_c = \sum_{j=1}^{12} \left[\hat{E}_j \right] \quad (7)$$

where:

\hat{E}_j = estimated annual number of emigrants from province j

\hat{E}_c = estimated annual number of emigrants from Canada

$E_{j,0-17}$ = number of emigrants from province j aged 0 to 17 who were eligible for Family Allowance

$P_{j,0-17}$ = number of children in province j who are eligible for Family Allowance

$P_{j,18+}$ = adult population of province j obtained by subtracting the number of children eligible for Family Allowance from the total estimated population

f_c = annual adjustment factor for estimating adult emigration from Canada, based on Revenue Canada tax files.

$E'_{c,18+}$ and $E'_{c,0-17}$ = estimated numbers of adult and child emigrants from Canada, based on Revenue Canada tax files.

$\hat{P}_{c,18+}$ and $\hat{P}_{c,0-17}$ = estimated June 1st population of adults and children for Canada, based on the component method.

The method of estimating the number of emigrants was modified in March 1989, affecting estimates after 1986. The new method combines counts by age of emigrants from Canada to the United States (from the U.S. Department of Justice, Immigration and Naturalization Service), and estimates of the numbers of emigrants from Canada to countries other than the U.S. based on Family Allowance files and an f_c factor calculated from immigration files (see Raby, Martel and Cartier 1989).

3. EVALUATION OF ESTIMATES OF THE COMPONENTS OF POPULATION CHANGE

Each component of population change (births, deaths, immigrants, emigrants and inter-provincial migrants) may contain a degree of bias and error. However, the data on births, deaths and immigration can be regarded as more accurate than the estimates of emigrants and inter-provincial migrants. In 1982, the methods of estimating emigrants and internal migration were thoroughly updated (see Statistics Canada 1987). These revised methods are evaluated below.

Table 1
Estimates of Emigrants by Different Methods, Canada, 1976-1981 and 1981-1986

Method	1976-81	1981-86
Residual*		
(a) Unadjusted	277,558	476,373
(b) Adjusted for Undercoverage	196,955 ¹	134,857 ¹
(c) Adjusted for Net Undercoverage	194,155 ²	218,148 ²
Revenue Canada Tax File	207,420	165,272
Family Allowance Method	278,624	235,481
Reverse Record Check	296,724	288,376

*Residual Method:
$$\text{Emigrants} = ([\text{Births} - \text{Deaths}] + [\text{Immigrants}]) - \text{Intercensal growth of population between time } t \text{ and } t + 5.$$

¹ The undercoverage rates were 2.04% for the 1976 Census, 2.01% for the 1981 Census, and 3.21% for the 1986 Census.
² The 1976, 1981 and 1986 Census net undercoverage rates were 1.53%, 1.51% and 2.40% respectively. They are estimated using the U.S. experience of overcoverage which is 25% of the undercoverage rate.

Source: Demography Division, Statistics Canada.

3.1 Emigration Data

Table 1 presents estimates of emigrants from Canada by using different methods and data sources for 1976-1981 and 1981-1986. For 1981-1986, the estimate using the residual method is considerably higher than the estimate based on the Family Allowance file. The residual method subtracts the population growth between 1981 and 1986, unadjusted for census undercoverage, from natural increase and immigration. Since births, deaths and immigration data are assumed to be accurate, the higher estimate by the residual method can be attributed to the difference in undercoverage rates for 1981 and 1986. After adjusting the 1981 and 1986 Census counts for undercoverage (2.01% and 3.21% respectively), the estimate by the residual method was found to be 134,857. This figure is lower than estimates obtained using both the Family Allowance file (235,481) and the Revenue Canada tax file (165,272).

This low estimate may result from different rates of overcoverage in the 1981 and 1986 Censuses. No estimate of overcoverage is calculated in the Reverse Record Check study, but the rate can be assumed to be similar to the U.S. Census rate which is 25% of the undercoverage rate. After adjusting the 1981 and 1986 Census counts for net coverage rates of 1.51% and 2.40% respectively, the residual estimate (218,148) was close to the Family Allowance-based estimate (235,481).

For 1976-1981, the estimating methods do not produce similar results. The number of emigrants estimated by the residual method adjusted for net undercoverage was 194,155, which is close to the estimate based on Revenue Canada tax files (207,420), but considerably lower than the Family Allowance method estimate (278,624) or the Reverse Record Check estimate (296,724).

One possible source of error in the current method is the f_c factors, which are adult-child emigrant ratios, estimating the number of emigrants aged 18+ from 1981-1986. These ratios were obtained from the emigration data provided by the Revenue Canada tax files.

Table 2 shows f_c values derived from different data sources. The f_c factors from the Revenue Canada tax files are less than unity and higher than unity from the three other data sources: interprovincial migration data from income tax files, immigration files, and data on Canadian emigrants to the United States. The estimates of emigrants from these sources are also higher than the Revenue Canada-based estimate.

Table 2
Estimates of Emigrants by Family Allowance Method Using Different Values
of f_c (Adult-Child Emigrant Ratios), 1981-1986

Data Source of f_c	Value of f_c Factor					Number of Emigrants
	1981-82	1982-83	1983-84	1984-85	1985-86	
1. Revenue Canada Tax Files	0.8698	0.8768	0.9052	0.8592	0.8592	235,481
2. Interprovincial Migration Data from Income Tax Files	1.0760	1.1000	1.0664	1.0290	1.0029	265,816
3. EIC Immigration Data	1.0801	1.0926	1.1723	1.1254	1.0694	275,762
4. Canadian Emigrants to the U.S.A.	1.2300	1.2774	1.3196	1.3745	1.4232	316,268

Source: Demography Division, Statistics Canada.

Each f_c factor source shows annual variations. The f_c factors for Canadians emigrating to the United States are particularly high, indicating that 23% to 42% more adults emigrated to the U.S. than did children. This is not surprising, as the southern American states have always been attractive to retirees. Hence the f_c factor based on U.S. data may not be suitable for estimating Canadian emigrants to countries other than the U.S.

Similarly, the f_c factors for interprovincial migration, based on the income tax file, suggest that adult migrants have exceeded child migrants by up to 10% from 1981 to 1986. However, the adult migrant group likely contains a high proportion of younger adults, who tend to move more often between provinces than other age groups. Hence this data source is also very specific and thus not suitable for computing the overall f_c factor.

According to some authors (Beaujot and Rappak 1988), emigrant and immigrant flow data are associated, making it possible to compute an f_c factor from the Employment and Immigration Canada (EIC) immigration file. f_c factors from the EIC immigration file are intermediate between those derived from interprovincial immigrant data and U.S. emigrant data. The figure based on the f_c factor from the immigration file (275,762) is higher than the official estimate of emigrants (235,481), but is close to that derived from the 1986 Reverse Record Check study (288,376). If the official estimate of the number of emigrants were increased to 275,762, the 1986 error of closure between the population estimate and census counts would be reduced from 0.95% to 0.79%.

In sum, for the 1981-86 period the estimates of emigrants seemed to be improved by taking f_c factors from the Canada Employment and Immigration (EIC) immigrant file rather than the Revenue Canada tax file.

Yet in March 1989, it was discovered that emigrant estimates based on Family Allowance files and an f_c factor derived from EIC immigration data were still too low after 1986. This seems to be a result of the high proportion (33%) of Canadian emigrants destined for the U.S. from 1981 to 1986, according to U.S. data.

An analysis was also made of a method combining U.S. Department of Justice, Immigration and Naturalization Service data on the numbers emigrating to the U.S. from Canada; child emigrant counts (ages 0-17) from Family Allowance files and an f_c factor obtained from the EIC immigration file for all countries other than the U.S. For 1981 to 1986, the estimated number of emigrants by this method was 285,413. This revised estimate is much closer to the Reverse Record Check study figure (288,376).

Table 3

Estimates of Net Interprovincial Migration from 1986 Census Data on Mobility, Family Allowance Files, Income Tax Files, and Residual Method, Canada, Provinces and Territories, 1981-1986

Geographic Area	1986 Census ¹	Family Allowance Files	Income Tax Files	Residual Method ²
CANADA	0	0	0	- 238,178
Nfld.	- 16,550	- 14,837	- 15,051	- 26,111
P.E.I.	1,540	293	751	- 509
N.S.	6,275	5,204	6,895	- 4,095
N.B.	- 1,370	- 2,239	- 65	- 11,212
Que.	- 63,295	- 76,040	- 81,254	- 167,286
Ont.	99,355	115,497	121,767	57,147
Man.	- 1,555	- 3,700	- 2,634	- 8,180
Sask.	- 2,820	- 668	- 2,974	- 13,564
Alta.	- 27,665	- 34,073	- 31,676	- 50,811
B.C.	9,500	13,289	7,382	- 12,418
Yukon	- 2,665	- 2,381	- 2,775	- 1,643
N.W.T.	- 755	- 345	- 366	504

¹ Population 5 years of age and over.

² The residual method for estimating net interprovincial migration is:

$$\text{Net Migration} = \text{Growth of Census Population between time } t \text{ and } t + 5 \\ - [(\text{Births} - \text{Deaths}) + (\text{Immigration} - \text{Emigration})].$$

Source: Demography Division, Statistics Canada.

3.2 Interprovincial Migration Data

To test the accuracy of estimates of interprovincial migration obtained from the Revenue Canada tax file, two evaluations were conducted: (i) a comparison of sets of interprovincial migration data derived from the Revenue Canada tax files and Family Allowance files; and (ii) a comparison of the errors of closure of population estimates for two sets of internal migration data.

Table 3 presents net interprovincial migration estimates derived from four sources: 1986 Census data on mobility; the Revenue Canada tax file; the Family Allowance file; and the residual-based net migration estimate. For all provinces, estimates of internal migration derived from the 1986 Census mobility data, the Revenue Canada tax file and Family Allowance files were consistent on the direction of net migration. All sources except the residual-based method show positive net migration for Prince Edward Island, Nova Scotia, Ontario and British Columbia. In other provinces, net migration was negative.

The estimates of net interprovincial migration from Family Allowance files and Revenue Canada tax files are not strictly comparable to the residual method. By definition, the sum of net interprovincial migration in Canada, should be zero. However, the sum produced using the residual method is about 238,000. In addition, the differences between the residual-based and the Revenue Canada/Family Allowance-based net interprovincial migration estimates are very high in Newfoundland, New Brunswick, Quebec, Ontario and Alberta.

The coefficient of variation (the ratio of the standard deviation of the average absolute error of closure for the provinces to the average absolute error of closure) was used to measure the relative accuracy of the internal migration estimates. The other estimates of the components of population change were assumed to be accurate. Statistically, a coefficient of variation of 20% to 30% is normally acceptable.

Table 4
Error of Closure Between Alternative Population Estimates and Census Counts
by Province and Territory 1971, 1976, 1981 and 1986

Geographic Area	Error of Closure ¹ (%)							
	1971		1976		1981		1986	
	Income Tax	FA	Income Tax	FA	Income Tax	FA	Income Tax	FA
Newfoundland	-2.08	-1.64	0.49	1.34	1.63	2.30	1.97	2.01
Prince Edward Island	-2.09	-2.01	0.17	2.11	-0.05	1.02	0.99	0.63
Nova Scotia	-1.68	-2.39	-0.20	1.18	0.30	0.40	1.24	1.04
New Brunswick	-1.93	-2.65	-1.29	1.81	0.13	0.54	1.58	1.04
Quebec	-0.33	-0.97	-0.05	-0.18	-0.30	-0.07	1.32	1.40
Ontario	0.11	0.99	0.15	0.16	0.64	0.37	0.72	0.65
Manitoba	0.29	0.38	-0.27	0.39	1.07	0.87	0.51	0.41
Saskatchewan	0.44	-0.33	0.45	0.37	-0.31	0.28	1.08	1.31
Alberta	-0.14	0.52	-1.07	-1.11	-2.39	-2.64	0.73	0.63
British Columbia	0.01	-1.34	0.28	-1.10	0.03	-0.07	0.59	0.79
Yukon	-5.36	-5.99	-0.87	3.79	-1.98	2.06	-4.78	-3.10
Northwest Territories	-2.12	2.64	-12.98	-3.39	-7.08	0.43	-1.44	-1.40
Average Absolute Error								
10 provinces	0.91	1.33	0.44	0.97	0.69	0.86	1.07	1.01
Provinces and Territories	1.38	1.82	1.52	1.41	1.33	0.92	1.41	1.22

Note: From 1976 to 1980, Revenue Canada data for children were available for age group 0-15 only. Therefore the $f_{(j,k)}$ factors were calculated using migrants aged 0-15 and 16+ instead of 0-17 and 18+.

¹ Error of closure is calculated using the following equation:

$$\text{Error of closure} = \left(\frac{\text{Estimate} - \text{Census count}}{\text{Census count}} \right) \times 100$$

Income Tax: Revenue Canada Income Tax File. FA: Family Allowance File.

Source: Estimates of interprovincial migration based on Family Allowance data, Demography Division, Statistics Canada.

Estimates of interprovincial migration based on tax data, Small Area and Administrative Development Division, Statistics Canada.

Table 5
Coefficients of Variation of the Average Absolute Error of Closure between the Population Estimates and Census Counts among Provinces ($n = 10$), by Source of Interprovincial Migration Estimates, 1966-1971, 1971-1976, 1976-1981 and 1981-1986

Period ($t, t + 5$)	Source	AAE ($t + 5$)	Standard Deviation	Coefficient of Variation (%)
		(1)	(2)	(3) = (2 ÷ 1) × 100
1966-1971	Income Tax	0.91	0.2863	31
	FA	1.33	0.2642	20
1971-1976	Income Tax	0.44	0.1317	30
	FA	0.97	0.2135	22
1976-1981	Income Tax	0.69	0.2463	36
	FA	0.86	0.2855	33
1981-1986	Income Tax	1.07	0.1496	14
	FA	1.01	0.1570	16

Note: AAE: Average absolute error of closure.

Income Tax: Revenue Canada Income Tax File.

FA: Family Allowance File.

Source: Demography Division, Statistics Canada.

However, one could argue that the coefficient of variation is not a good indicator of the quality of internal migration data. For example, a set of estimates with an absolute error of closure of 10% for every province would give a coefficient of variation of zeros and consequently would be preferable to a set of estimates with closure errors ranging between -1.0% and 1.0%. For cases like this, a quality measure that takes into account the size of the absolute error of closure as well as the standard deviation of absolute closure errors is clearly required. However, the likelihood of the provinces having the same absolute error of closure is extremely low (see Table 5), hence, the application of the coefficient of variation in this paper seemed to be valid.

Table 5 shows the coefficient of variation (computed from figures in Table 4) for population estimates based on two sets of internal migration estimates and the census counts for 1971, 1976, 1981 and 1986. Before 1976, the coefficients of variation for migration data from tax files were 50% higher for data from the Family Allowance file. This was expected, since the method for estimating migration from tax files was in the developmental stage. Furthermore, in estimating the number of interprovincial migrants, the f_j factor (adult to child migration rates) was based on Census mobility data, an approach found to be less satisfactory than the current method. However, for 1976-1981 and 1981-1986, the gap in the coefficient of variation between the tax and Family Allowance migration data narrowed considerably.

The tax-based migration data coefficient of variation was 9% higher in 1981 and 12% lower in 1986 than the coefficient of variation based on the Family Allowance file. Hence, the two sets of data are comparable, producing similar provincial estimates and errors of closure with the same level of variation among provinces. Since the coefficient of variation for each set is under 20%, they provide acceptable data on internal migration.

In conclusion, estimates of interprovincial migration from the Revenue Canada tax files for 1981-1986 are consistent with estimates from the Family Allowance file. By province, they yield small variations in the errors of closure.

4. CONCLUSION AND DISCUSSION

The Family Allowance files and Revenue Canada tax files play important roles in providing consistent emigration and internal migration estimates for Canada, and for the provinces and territories. For 1981 to 1986, estimates of emigrants and interprovincial migrants obtained from these files are acceptable for estimating total population.

Nationally the error of closure (the difference between the population estimates and census counts) for 1986 was higher than for the census years 1971, 1976 and 1981. In addition, the errors of closure by province in 1986 were positively biased, indicating that in all provinces the estimates were higher than census counts.

These discrepancies are largely a result of differences in coverage of the 1981 Census population, which was used as the bench-mark, and coverage of the 1986 Census population. The Reverse Record Check estimate of the 1981 undercoverage rate for Canada was 2.01%. The estimate for the 1986 Census was considerably higher, 3.21%.

Errors in the estimates of the other components of change may also partly account for the discrepancies.

ACKNOWLEDGEMENT

We are grateful to the reviewers of this article for their comments and suggestions.

REFERENCES

- BEAUJOT, R., and RAPPAK, J.P. (1988). *Emigration from Canada: its Importance and Interpretation*. Ottawa: Employment and Immigration Canada.
- NORRIS, D., BRITTON, M., and VERMA, R.B.P. (1982). The use of administrative records for estimating migration and population. *Statistics of Income and Related Administrative Record Research: 1982*, Washington, D.C.: Department of the Treasury, Internal Revenue Service.
- NORRIS, D., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Technical Report, Small Area and Administrative Data Development Division, Statistics Canada.
- RABY, R., MARTEL, J., and CARTIER, G. (1989). Issues in the current postcensal population estimates. Paper presented at the Federal-Provincial Committee on Demography, Ottawa.
- STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue 91-528E, Statistics Canada.
- VERMA, R.B.P., and PARENT, P. (1985). An overview of the strengths and weaknesses of the selected administrative data files. *Survey Methodology*, 11, 171-179.

Confidence Intervals for Postcensal Population Estimates: A Case Study for Local Areas

DAVID A. SWANSON¹

ABSTRACT

This paper presents a technique for developing appropriate confidence intervals around postcensal population estimates using a modification of the ratio-correlation method termed the rank-order procedure. It is shown that the Wilcoxon test can be used to decide if a given ratio-correlation model is stable over time. If stability is indicated, then the confidence intervals associated with the data used in model construction are appropriate for postcensal estimates. If stability is not indicated, the confidence intervals associated with the data used in model construction are not appropriate, and, moreover, likely to overstate the precision of postcensal estimates. Given instability, it is shown that confidence intervals appropriate for postcensal estimates can be derived using the rank-order procedure. An empirical example is provided using county population estimates for Washington state.

KEY WORDS: Population estimation; Confidence intervals; Ratio-correlation regression.

1. INTRODUCTION

A method of generating confidence intervals for postcensal estimates was not available until Espenshade and Tayman (1982) introduced a time-series regression estimation technique utilizing age-specific postcensal death rates. The Espenshade-Tayman technique represents an important breakthrough in estimation technology; however, like most breakthroughs it has limitations, of which two are notable:

1. The technique is likely to be unsatisfactory at the subprovincial or substate level (Espenshade and Tayman 1982); and
2. It is a major departure from the standard regression technique used in Canada and the United States for estimating county-equivalent populations, namely, ratio-correlation. This departure is a particularly salient issue in terms of data requirements and the experience of people responsible for making county-equivalent and other subprovincial level population estimates. (Statistics Canada 1987). The term "county equivalent" is defined as a Census Division in Canada (Statistics Canada 1987) and as a county in nearly all U.S. states; notable exceptions in the U.S. include Alaska, in which county-equivalents are Census Areas, Louisiana, where Parishes function as counties, and Virginia, in which "independent cities" are included as county-equivalents.

This paper presents a means of developing confidence intervals for postcensal county-equivalent populations using the rank-order procedure, a modification of the ratio-correlation method introduced by Swanson (1980) that exploits causal modeling concepts to take into account postcensal structural changes in a given ratio-correlation model.

There are three issues relevant to the development of confidence intervals made using the ratio-correlation method. The first has to do with model stability over time. If the structure of associations among model variables is invariant over time, then the confidence intervals

¹ David A. Swanson, Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447, U.S.A.

constructed in regard to the model data set will apply to the population estimates generated by the model from the estimation data set. Although it has been consistently documented that it is not prudent to assume model invariance (D’Allesandro and Tayman 1980; Ericksen 1973, 1974; Mandell and Tayman 1982; Namboodiri 1972; O’Hare 1976, 1980; Smith and Mandell 1984; Spar and Martin 1979; Swanson 1980; Swanson and Prevost 1986; Swanson and Tedrow 1984; Tayman and Schafer 1982; Verma *et al.* 1983), it would be useful to have a testing procedure for stability. This leads to the second issue, namely, the use of a statistical test. If the test indicates that stability can not be assumed, and yet confidence intervals associated with, say, a model constructed using 1960-70 data, are applied to estimates generated for, say, 1979, they are likely to overstate the level of precision in the 1979 estimates. Thus, the third issue is the need for a procedure that will generate appropriate confidence intervals.

In the report that follows, a description of ratio-correlation is provided along with the modification that forms the basis for developing appropriate confidence intervals. Next, the logic for developing these confidence intervals is formally described, followed by an empirical example showing both the test for instability and the generation of both “inappropriate” and “appropriate” confidence intervals.

2. METHODOLOGY FOR POPULATION ESTIMATION

Ratio-correlation is a regression method designed to measure the temporal change in county-equivalent population proportions using observed temporal change in proportions of symptomatic indicators such as registered voters, covered employment and public school enrollment. The temporal change is measured by simply taking a ratio of proportions at two points in time.

Since enumerated population numbers for all county-equivalents are available only from the federal census, a ratio-correlation regression model is always constructed using two points in time separated by a regular number of years. It is formally described as

$$R_{it} = a_o + \sum_{j=1}^k (b_j) (X_i)_{jt} + \epsilon$$

where

- a_o = the intercept term to be estimated
- b_j = the regression coefficient to be estimated
- ϵ = the error term
- j = symptomatic indicator, $(1 \leq j \leq k)$
- i = county-equivalent $(1 \leq i \leq n)$
- t = the year of the most recent census

and

$$R_{it} = \left[\frac{P_{i,t}}{\sum P_{i,t}} \right] \div \left[\frac{P_{i,t-z}}{\sum P_{i,t-z}} \right] \tag{1.A}$$

$$(X_i)_{t,j} = \left[\frac{S_{i,t}}{\sum S_{i,t}} \right] \div \left[\frac{S_{i,t-z}}{\sum S_{i,t-z}} \right]_j \tag{1.B}$$

where

Z = the number of years between each census

P = Population

S = Symptomatic Indicator

Once a model is constructed, it is used to develop a postcensal estimate for time $t + x$ by substituting $(S_{i,t+x} / \sum S_{i,t+x})_j$ into the numerator of the right-hand side of equation [1.B] while $(S_{i,t} / \sum S_{i,t})_j$ is substituted into the denominator of the right-hand side of equation [1.B]. This means that once $\hat{R}_{i,t+x}$ is obtained, an actual population for area i at time $= t + x$ is developed by introducing an independently estimated total population, P_{t+x} , into equation [1.A] and algebraically solving equation [1.A] for $P_{i,t+x}$. Since $\sum \hat{P}_{i,t+x}$ does not usually equal the independently derived total, P_{t+x} , an adjustment is made to force the summed population figures to the independently estimated total.

One limitation of ratio-correlation is that its structure is invariant over time, which is why the rank order procedure was introduced by Swanson (1980). The rank-order procedure is based on the fact that information contained in the zero-order correlations found in an estimation data set can be exploited due to work by Land (1969, Chapter IV); work that is based on the fundamental theorem underlying path analysis as developed by Wright (1921). It involves a theoretical reversal of the dependent variable in the regression model, the population variable, as an unmeasured, causally prior variable and a just-identified structure – a minimum of three predictor variables (in the regression model), the covariance of which is assumed to be due to the fact that they are all causally related to the population variable.

3. METHODOLOGY FOR CONFIDENCE INTERVALS ESTIMATION

If the relationships found among the variables in the model data set remain stable over time (as shown through the rank-order procedure) then the same relationships should be found among the variables in the estimation data set. This stability would indicate that the S.E.E. associated with the model data set is appropriate for generating confidence intervals for the estimation data set. However, if stability does not exist, then the S.E.E. associated with the model data set is not appropriate, and may, in fact, generate confidence intervals that overstate the precision of postcensal estimates. These considerations lead to the question of determining stability through statistical inference.

In answering the question just posed, consider that we are examining related pairs of variables. This implies that the Wilcoxon matched-pairs signed rank test could be used (Mosteller and Rourke 1973). In using this test, the null hypothesis is that there are no differences between the population estimates (scores) produced by the unmodified and modified regresion models.

The key to developing confidence intervals for postcensal county equivalent population estimates is found in the fact that the rank-order procedure generates a set of regression coefficients for the estimation data set. From these coefficients, estimates of R^2 and the S.E.E. for the estimation data set can be developed, and the estimated S.E.E. leads directly to the

development of confidence intervals. First, recall that the coefficient of multiple determination, R^2 , is simply the sum of the products of each zero-order correlation between an independent variable and the dependent variable, and the standardized regression coefficient for each independent variable (Hayes 1973), so that S.E.E. is (Hayes 1973)

$$\text{S.E.E.} = \left[\frac{(n) (S_y^2) (1 - R^2)}{n - 2} \right]^{1/2}$$

where

n = number of cases (county-equivalents)

S_y^2 = variance of the dependent variable

R^2 = coefficient of multiple determination

The formula for generating a confidence interval around a given estimated value for a point on a (population) regression line is provided by Kmenta (1971)

$$Y_i \pm (t_{n-2, \alpha/2}) (\text{S.E.E.})$$

An important point to realize is that the confidence interval is not directly generated for a population estimate, rather it is for the estimated ratio of proportions, or R_{it+x} . However, as shown by Espenshade and Tayman (1982), a confidence interval around one variable can be translated for another variable algebraically substituted for the first. Thus, by finding the lower and upper confidence boundaries of R_{it+x} , these lower and upper confidence boundaries can be translated into the population values:

$$\begin{aligned} & (R_{it+x}) \pm (t_{n-2, \alpha/2}) (\text{S.E.E.}) \\ &= \left[\frac{P_{it+x}}{\sum P_{it+x}} \right] \div \left[\frac{P_{it}}{\sum P_{it}} \right] \pm (t_{n-2, \alpha/2}) (\text{S.E.E.}) \end{aligned}$$

which leads to

$$\begin{aligned} \text{L.L. } (\hat{P}_{it+x}) &= \\ & \left[\frac{P_{it}}{\sum P_{it}} \right] (\sum P_{it+x}) \left[(\hat{R}_{it+x}) - (t_{n-2, \alpha/2}) (\text{S.}\hat{\text{E.}}\text{E.}) \right] \end{aligned}$$

and

$$\begin{aligned} \text{U.L. } (\hat{P}_{it+x}) &= \\ & \left[\frac{P_{it}}{\sum P_{it}} \right] (\sum P_{it+x}) \left[(\hat{R}_{it+x}) + (t_{n-2, \alpha/2}) (\text{S.}\hat{\text{E.}}\text{E.}) \right] \end{aligned}$$

4. EMPIRICAL STUDY

Table 1.A in Swanson (1980) gives the zero-order correlations relating to a ratio-correlation model for estimating county civilian populations under sixty-five years from employment, voters, and grades 1-8 enrollment for the state of Washington, for the period 1950-1960. Characteristics of the model constructed from these data are given in Table 1.B. while Tables 2.A and 2.B provide similar results for the 1960-1970 period as found in Swanson (1980). This latter set forms the estimation data over which the procedure will be described.

Although full knowledge of the estimation data set is available, the procedure is used as if this were not the case. Of course, what is known in any estimation problem is the zero-order correlation matrix for the independent variables, which is used in conjunction with the fundamental theorem of path analysis to estimate the coefficients for the modified model. Using the complete rank-order procedure, the modified model (Swanson 1980) is:

$$Y = 0.046618 + 0.066786X_1 + 0.50727X_2 + 0.38736X_3.$$

Estimates for 1970 of the county civilian population under sixty-five years of age (adjusted to the independently estimated state total) resulting from the preceding modified model are presented in Table 1 along with the actual enumerated populations.

The Wilcoxon test was conducted for the Washington data using the procedure in the SPSSx NPAR Tests command (SPSS 1986). To save space, the unmodified and modified population estimates are not presented. They can be found in Table 3 of Swanson (1980). Under the null hypothesis, the probability of obtaining $Z = -3.2096$ is 0.0013. Thus, the null hypothesis is rejected and it is assumed that instability exists for Washington counties in going from the model constructed using 1960/1950 data to the true unknown model associated with 1970/1960 data.

As a note of interest, the Chow test (Chow 1960) validated the results of the Wilcoxon test by showing that the difference between the "true" 1970-1960 ratio-correlation model and the 1960/1950 ratio-correlation model was statistically significant.

Had the results of the Wilcoxon test led us not to reject the null hypothesis, we would have used the unmodified coefficients from the 1960/1950 model data set to generate 1970 population estimates for Washington counties. Further, the S.E.E. for this same model (0.05022) would have been used to generate confidence intervals for the 1970 estimates. However, the results of the Wilcoxon test led us to reject the null hypothesis in this case. This indicates the modified coefficients developed using the rank-order procedure should be used in lieu of the unmodified model. Further, it indicates the need for a revised S.E.E., one that is not likely to overstate the precision of the 1970 estimates.

Using the estimated values found in the 1970 example data for Washington state (Swanson 1980) we find

$$\hat{R}^2 = (0.07533) (0.75290) + (0.47085) (0.92146) + (0.49481) (0.88082) = 0.926$$

and

$$\begin{aligned} (\text{S.}\hat{\text{E.}}\text{E.}) &= \left[\frac{(39) (0.2145)^2 (1 - 0.926)}{39-2} \right]^{1/2} \\ &= 0.0599 \end{aligned}$$

Table 1
 90% Confidence Interval for the Estimated Civilian Population
 Under Sixty-Five Years by County,
 State of Washington 1970

County	Enumerated Population	Lower Limit	Estimated Population	Upper Limit	90% Confidence Interval (in percent)
Adams	11102	10335	11458	12581	± 9.80
Asotin	11862	10469	11814	13154	± 11.38
Benton	63144	60405	67511	74616	± 10.53
Chelan	35862	31733	36177	40620	± 12.28
Clallam	30023	28063	31294	34525	± 10.32
Clark	116663	101183	111437	121690	± 9.20
Columbia	3771	3683	4161	4639	± 11.49
Cowlitz	62586	55170	61581	67992	± 10.41
Douglas	15287	14569	16252	17935	± 10.36
Ferry	3336	2963	3397	3831	± 12.78
Franklin	23983	21960	24631	27302	± 10.84
Garfield	2546	2447	2761	3075	± 11.37
Grant	38921	37561	42606	47651	± 11.84
Grays Harbor	52583	46294	52114	57935	± 11.17
Island	20589	20512	22148	24040	± 7.39
Jefferson	9235	8440	9473	10506	± 10.90
King	1054271	935664	1037937	1140203	± 9.85
Kitsap	86529	77022	85821	94619	± 10.25
Kittitas	22764	17649	19863	22077	± 11.15
Klickitat	10729	10440	11923	13406	± 12.44
Lewis	39265	35747	40122	44497	± 10.90
Lincoln	8168	7939	9107	10275	± 12.83
Mason	18411	16057	17827	19596	± 9.93
Okanogan	22952	21002	23795	25688	± 10.97
Pacific	13310	11270	12795	14320	± 11.92
Pend Oreille	5185	5147	5893	6639	± 12.86
Pierce	339048	314272	346728	379184	± 9.36
San Juan	3089	2636	2918	3201	± 9.66
Skagit	45703	43255	48758	54261	± 11.29
Skamania	5330	4787	5358	5929	± 10.66
Snohomish	245193	213164	231996	250827	± 8.12
Spokane	251057	227372	256723	286072	± 11.43
Stevens	15178	13869	15780	17692	± 12.11
Thurston	68719	63644	69540	75436	± 8.48
Wahkiakum	3137	3033	3397	3761	± 10.72
Walla Walla	36608	33727	38271	42812	± 11.87
Whatcom	72111	63218	70670	78122	± 10.54
Whitman	34843	28960	32409	35858	± 10.64
Yakima	128960	120347	136203	152219	± 11.69

Note, that from Table 2 in Swanson (1980), the actual R^2 and S.E.E. values are 0.878 and 0.05077, respectively. In comparison with the actual S.E.E. of 0.05077, the estimated S.E.E. is higher. This is appropriate given that we are more uncertain about the precision of estimates generated by the rank-order procedure than we would be about the precision associated with the “true” model, if in fact, the true model was obtainable. With the rank-order procedure, we can now generate a confidence band from the following formula:

$$Y_i \pm (t_{37,\alpha/2}) (0.0599)$$

In Table 1 an empirical example using a 90% confidence interval is given for the 1970 estimated county population figures presented also in Table 1. Here, the 90% confidence interval is given by:

$$\left[\frac{P_{i1960}}{2522141} \right] (3032053) \left[(\hat{R}_{i1970}) \pm (1.69) (0.0599) \right]$$

In examining the confidence intervals given in Table 1 in combination with the enumerated populations provided, it is found that in only one county (Kittitas) is the enumerated population outside of the 90% confidence interval. In this instance, the enumerated population exceeds the upper limit by 687 people. At a 90% level of confidence, the intervals are fairly wide, with a mean of 10.81, a minimum of ± 7.39 percent for Island county and a maximum of ± 12.83 percent in Lincoln County. Compare these with the mean of the absolute percent errors associated with the 1970 estimates, which is 4.89 (Swanson 1980). This comparison suggests that the 90% level generates intervals that are too broad for practical use. Given this, it is of interest to consider which level of confidence would be more appropriate. It is also of interest to consider the effect of using the unmodified S.E.E. (0.05022) from the 1960/1950 model. We would expect that the confidence intervals generated by the unmodified model would be too optimistic. That is, at a given level of confidence, there would be fewer than expected counties for which the interval encompassed the actual population. To explore these issues, Table 2 was constructed.

In Table 2, two distinct sets of information are provided. For both sets, however, a comparison is made between the unmodified and modified estimates and their associated confidence intervals. In regard to the issue of expecting optimistic confidence intervals for the 1970 estimates generated by the unmodified model, Table 2 indicates that at varying levels of confidence ranging from 90% down to 50%, the intervals are, indeed, optimistic in that for only two of the six levels examined are the expected number of county estimates within the specified level of precision. At the 80% level, for example, only 28 (72 percent) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates; at the 60% level, only 22 (56%) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates.

The second aspect of Table 2 is the mean interval associated with a given level of confidence. At the 90% level, the mean of the intervals associated with the unmodified model is 9.10 percent; for the modified model it is 10.81 percent. At the 50% level, the means are 3.66% and 4.35%, respectively. Thus, it is clear that the 60% and 50% levels of confidence generate a mean interval that is more in line with the mean absolute percent error, which is 4.88 for the modified model.

Table 2
Number (%) of Counties in Which Actual 1970 Population
was Inside the Confidence Interval

Level of Confidence	Unmodified S.E.E. (0.05022)	Modified S.E.E. (0.0599)
90%	35 (89.7%)	38 (97.4%)
80%	28 (71.8%)	33 (84.6%)
70%	24 (61.5%)	29 (80.6%)
66.66%	24 (61.5%)	26 (66.66%)
60%	22 (56.4%)	23 (59.0%)
50%	20 (51.3%)	22 (56.4%)
Mean Interval (in percent)		
	Unmodified S.E.E. (0.05022)	Modified S.E.E. (0.0599)
90%	9.10	10.81
80%	7.02	8.38
70%	5.66	6.75
66.66%	5.59	6.40
60%	4.59	5.47
50%	3.66	4.35

In examining the issue of confidence intervals, it appears that a procedure is needed for generating confidence intervals that are not misleading in terms of the precision of postcensal county-equivalent population estimates. However, guidance is also needed on selecting a given level of confidence that is appropriate for the estimates. Of interest in this regard is the work of Stoto (1983) on empirical confidence intervals for population projections. One of Stoto's (1983:18) findings is the high and low population projections produced for the United States by the Bureau of the Census (1977) correspond to a 66.66% confidence interval. It may be the case that for county-equivalent postcensal populations, that the 66.66% confidence level is also appropriate, although in this test this level of confidence generates a mean interval of 6.4 percent for the modified estimates, which is somewhat above their mean percent error (4.9). Another consideration is the length of time between the year for which a postcensal estimate is desired and the preceding census. In the example, the maximum period of postcensal time in the United States was used, 10 years. For each county, we have, in essence, a situation in which maximum uncertainty exists in regard to estimates. From this perspective, the relatively wide interval generated for each county at a 90 percent level of confidence is appropriate. We would expect that structural model changes occur relative to time. Hence, a narrower band would likely be generated in the first year following the end-census year of model construction than in the second year; and so on through the intercensal period.

5. CONCLUSION

At this point it should be clear that the rank-order procedure is not being presented as a fully-validated technique for constructing confidence intervals around postcensal county-equivalent population estimates. However, it appears to offer a reasonable starting point. Even with its limitations, the use of the Wilcoxon test and the confidence intervals developed using the rank-order procedure appears capable of providing benefits to those responsible for making such postcensal population estimates. In the first place, as noted by Espenshade and Tayman (1983), it is important to provide the users of postcensal population estimates some notion of their accuracy as do both the Wilcoxon test and the confidence intervals. Second, with the selection of appropriate confidence intervals, a formal means is available for resolving disputes over the population of a given county-equivalent by using hypothesis testing procedures. Third, S.E.E. can be used as a basis for selecting one model over another. This means that a set of different ratio-correlation models could be considered for any given postcensal estimation year and, further, that a formal criterion is available for selecting one model over another. This feature could be useful in the event that the ratio-correlation estimates generated by a federal, provincial or state demographic center, are challenged in a given postcensal year, an event that has become more frequent, especially in the U.S. (D'Allesandro 1987).

ACKNOWLEDGMENTS

The Author is grateful to Jeff Tayman, anonymous referees, and the editorial staff for comments and suggestions. Peggy Jobe typed a draft of this paper and Carey Taylor typed the final version.

REFERENCES

- CHOW, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28, 591-605.
- D'ALLESANDRO, F. (1987). Should applied demographers take out liability insurance? Paper presented at the Annual Meeting of The Population Association of America.
- D'ALLESANDRO, F., and TAYMAN, J. (1980). Ridge regression for population estimation: Some insights and clarification. *Staff Document No. 56*. Office of Financial Management, State of Washington: Olympia, Washington.
- DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis, 2nd Edition*. New York: Wiley.
- ERICKSEN, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.
- ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.
- ESPENSHADE, T.J., and TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.
- HAYS, W.L. (1973). *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston.
- KMENTA, J. (1971). *Elements of Econometrics*. New York: Macmillan.
- LAND, K.C. (1969). Explorations in mathematical sociology. Unpublished Ph.D. dissertation. University of Texas, Austin.

- MANDELL, M., and TAYMAN, J. (1982). Measuring temporal stability in regression models of population estimation. *Demography*, 19, 1351-46.
- MOSTELLER, F., and ROURKE, R. (1973). *Sturdy Statistics*. Reading, Massachusetts: Addison-Wesley.
- NAMBOODIRI, N.K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography*, 9, 443-453.
- O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.
- O'HARE, W. (1980). A note on the use of regression methods in population estimates. *Demography*, 17, 341-343.
- SMITH, S., and MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 99, 292-289.
- SPAR, M., and MARTIN, J. (1979). Refinements to regression-based estimates of postcensal population characteristics. *Review of Public Data Use*, 7, 16-22.
- SPSS, Inc. (1986). *SPSSx User's Guide*. Chicago: SPSS, Inc.
- STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E, Statistics Canada.
- STOTO, M.A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78, 13-20.
- SWANSON, D. (1980). Improving accuracy in multiple regression estimates of population using principles from causal modeling. *Demography*, 17, 413-427.
- SWANSON, D., and PREVOST, R. (1986). Identifying extreme errors in ratio-correlation estimates of population. Presented at the Annual Meeting of the Population Association of America.
- SWANSON, D., and TEDROW, L. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography*, 21, 373-381.
- TAYMAN, J., and SCHAFER, E. (1985). The impact of coefficient drift and measurement error on the accuracy of ratio-correlation population estimates. *The Review of Regional Studies*, 15, 3-10.
- U.S. BUREAU OF THE CENSUS (1977). Projections of the Population of the United States, 1977 to 2050. *Current Population Reports. Series P-25 No. 704*. Washington, D.C.: U.S. Government Printing Office.
- VERMA, R.V.P., BASAVARAJAPPA, K.G., and BENDER, R.K. (1983). The regression estimates of population for subprovincial areas in Canada. *Survey Methodology*, 9, 219-240.
- WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.

ACKNOWLEDGEMENTS

The Survey Methodology journal wishes to thank the following persons who have served as referees, sometimes more than once, during 1989.

- | | |
|---|--|
| M.G. Arellano, <i>University of California at San Francisco</i> | P. Lavallée, <i>Statistics Canada</i> |
| T.R. Balakrishnan, <i>University of Western Ontario</i> | B. Lefrançois, <i>Statistics Canada</i> |
| M. Bankier, <i>Statistics Canada</i> | R.J.A. Little, <i>University of California at Los Angeles</i> |
| K. G. Basavarajappa, <i>Statistics Canada</i> | L. Mach, <i>Statistics Canada</i> |
| D.R. Bellhouse, <i>University of Western Ontario</i> | E. Martin, <i>U.S. Bureau of the Census</i> |
| L. Biggeri, <i>University of Florence</i> | K. Namboodiri, <i>Ohio State University</i> |
| D. Binder, <i>Statistics Canada</i> | M. Otto, <i>U.S. Bureau of the Census</i> |
| L. Blais, <i>Statistics Canada</i> | J.N.K. Rao, <i>Carleton University</i> |
| M. Brick, <i>Westat</i> | D.B. Rubin, <i>Harvard University</i> |
| P. Cholette, <i>Statistics Canada</i> | I. Sande, <i>Bell Communications Research</i> |
| G.H. Choudhry, <i>Statistics Canada</i> | C.E. Särndal, <i>University of Montreal</i> |
| T.C. Chua, <i>National University of Singapore</i> | J. Schafer, <i>Harvard University</i> |
| C.D. Cowan, <i>National Center for Education Statistics</i> | N. Schenker, <i>University of California at Los Angeles</i> |
| E.B. Dagum, <i>Statistics Canada</i> | F. Scheuren, <i>U. S. Internal Revenue Service</i> |
| W.A. Fuller, <i>Iowa State University</i> | G.A. Shababb, <i>NPD/Nielsen, Inc.</i> |
| J. Gambino, <i>Statistics Canada</i> | A.C. Singh, <i>Statistics Canada</i> |
| J. Gentleman, <i>Statistics Canada</i> | C.J. Skinner, <i>University of Southampton</i> |
| M. Gonzalez, <i>U. S. Office of Management and Budget</i> | J. Smith, <i>Statistics Canada</i> |
| G.B. Gray, <i>Statistics Canada</i> | B.D. Spencer, <i>Northwestern University</i> |
| M. Hidioglou, <i>Statistics Canada</i> | K.P. Srinath, <i>Statistics Canada</i> |
| S. Hillmer, <i>University of Kansas</i> | S. Sudman, <i>University of Illinois</i> |
| D. Holt, <i>University of Southampton</i> | A. Sunter, <i>A.B. Sunter Research Design & Analysis, Inc.</i> |
| P. Hoyt, <i>Statistics Canada</i> | J.-L. Tambay, <i>Statistics Canada</i> |
| T.B. Jabine, <i>Statistical Consultant</i> | A. Théberge, <i>Statistics Canada</i> |
| G. Kalton, <i>University of Michigan</i> | V. Tremblay, <i>Statplus</i> |
| P.S. Kott, <i>U.S. Department of Agriculture</i> | A.R. Tupek, <i>U.S. Bureau of Labor Statistics</i> |
| P. Krishnan, <i>University of Alberta</i> | R. Verma, <i>Statistics Canada</i> |
| S. Kumar, <i>Statistics Canada</i> | G. Werking, <i>U.S. Bureau of Labor Statistics</i> |
| N.M. Lalu, <i>University of Alberta</i> | W.E. Winkler, <i>U.S. Bureau of the Census</i> |
| E. Langlet, <i>Statistics Canada</i> | K. Wolter, <i>A. C. Nielsen</i> |

Acknowledgements are also due to those who assisted during the production of the 1989 issues: C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services). Finally we wish to acknowledge J. Clarke, J. Dufresne, M. Kent, C. Lacroix, C. Larabie and D. Lemire for their support with coordination, typing and copy editing.

SPECIAL OFFER

Copies of the proceedings of recent Statistics Canada symposia are still available, and may be purchased at a nominal cost. These are:

Symposium 87: *Statistical Uses of Administrative Data*. Two volumes, English and French. Regular price, each: \$35.

These are now available at \$10 each or \$12 for both languages.

Symposium 88: *The Impact of High Technology on Survey Taking*. Bilingual, English and French. Regular price, each: \$20.

These are now available at \$10.

Cheques or money orders should be made payable to:

“The Receiver General for Canada”.

Requests for these volumes, along with cheque or money order should be sent to:

Production Manager
Survey Methodology
Statistics Canada
4-C2 Jean Talon Building
Tunney's Pasture
Ottawa, Ontario
Canada K1A 0T6

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, l).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.
4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

OFFRE SPÉCIALE

Des copies des actes des plus récents symposiums de Statistique Canada sont toujours disponibles, et peuvent être obtenus à un coût minime. Ces actes sont:

Symposium 87: *Les Utilisations Statistiques des Données Administratives*. Deux volumes, français et anglais. Prix régulier: 35\$ chacun.

Ils sont maintenant disponibles pour 10\$ chacun, ou 12\$ pour un exemplaire dans chacune des deux langues.

Symposium 88: *Les Répercussions de la Technologie de Pointe sur les Enquêtes*. Bilingue, français et anglais. Prix régulier: 20\$ chacun.

Ils sont maintenant disponibles pour 10\$ chacun.

Les chèques ou virements bancaires doivent être libellés à l'ordre du:

„Receveur Général du Canada”.

Les demandes pour ces volumes, accompagnées d'un chèque ou d'un virement bancaire, doivent être envoyées à:

Directeur de la Production
Techniques d'Enquête
Statistique Canada
4-C2 Immeuble Jean Talon
Parc Tunney
Ottawa, Ontario
Canada K1A 0T6

REMERCIEMENTS

La revue Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1989.

M.G. Arellano, *University of California at San Francisco*
T.R. BalaKrishnan, *University of Western Ontario*
M. Bankier, *Statistique Canada*
K.G. Basavarajappa, *Statistique Canada*
D.R. Bellhouse, *University of Western Ontario*
L. Biggert, *Université de Florence*
D. Binder, *Statistique Canada*
L. Blais, *Statistique Canada*
M. Brick, *Westat*
P. Cholette, *Statistique Canada*
G.H. Choudhry, *Statistique Canada*
T.C. Chua, *National University of Singapore*
C.D. Cowan, *National Center for Education Statistique*
E.B. Dagum, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
J. Gentleman, *Statistique Canada*
M. Gonzalez, *U. S. Office of Management and Budget*
G.B. Gray, *Statistique Canada*
M. Hidiroglou, *Statistique Canada*
S. Hillmer, *University of Kansas*
D. Holt, *University of Southampton*
P. Hoyt, *Statistique Canada*
T.B. Jabine, *Expert-conseil en statistique*
G. Kalton, *University of Michigan*
P.S. Kott, *U.S. Department of Agriculture*
P. Krishnan, *University of Alberta*
S. Kumar, *Statistique Canada*
N.M. Laju, *University of Alberta*
E. Langlet, *Statistique Canada*
K. Wolter, A. C. Nielsen
P. Lavallée, *Statistique Canada*
B. Lefrançois, *Statistique Canada*
R.J.A. Little, *University of California at Los Angeles*
L. Mach, *Statistique Canada*
E. Martin, *U.S. Bureau of the Census*
K. Namboodiri, *Ohio State University*
M. Otto, *U.S. Bureau of the Census*
J.N.K. Rao, *Carleton University*
D.B. Rubin, *Harvard University*
I. Sande, *Bell Communications Research*
C.E. Särndal, *Université de Montréal*
J. Schaffer, *Harvard University*
N. Schenker, *University of California at Los Angeles*
F. Scheuren, *U. S. Internal Revenue Service*
G.A. Shababb, *NPD/Nielsen, Inc.*
A.C. Singh, *Statistique Canada*
C.J. Skinner, *University of Southampton*
J. Smith, *Statistique Canada*
B.D. Spencer, *Northwestern University*
K.P. Srinath, *Statistique Canada*
S. Sudman, *University of Illinois*
A. Sunter, A.B. *Sunter Research Design & Analysis, Inc.*
J.-L. Tambay, *Statistique Canada*
A. Thèberge, *Statistique Canada*
V. Tremblay, *Statplus*
A.R. Tupek, *U.S. Bureau of Labor Statistique*
R. Verma, *Statistique Canada*
G. Werking, *U.S. Bureau of Labor Statistique*
W.E. Winkler, *U.S. Bureau of the Census*
E. Langlet, *Statistique Canada*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1989: C. VanBastelaar (Photocomposition), G. Gaullin (Services aux auteurs) et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à J. Clarke, J. Dufresne, M. Kent, C. Lacroix, C. Larabie et D. Lemire pour leur apport à la coordination, la dactylographie et la rédaction.

- BRICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.
- ESPENSHADE, T.J., et TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.
- HAYS, W.L. (1973). *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston.
- KMENTA, J. (1971). *Elements of Econometrics*. New York: Macmillan.
- LAND, K.C. (1969). Explorations in mathematical sociology. Thèse de doctorat. University of Texas, Austin.
- MANDELL, M., et TAYMAN, J. (1982). Measuring temporal stability in regression models of population estimation. *Demography*, 19, 1351-46.
- MOSTELER, F., et ROURKE, R. (1973). *Sturdy Statistics*. Reading, Massachusetts: Addison-Wesley.
- NAMBOODIRI, N.K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography*, 9, 443-453.
- O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.
- O'HARE, W. (1980). A note on the use of regression methods in population estimates. *Demography*, 17, 341-343.
- SMITH, S., et MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 99, 292-289.
- SPAR, M., et MARTIN, J. (1979). Refinements to regression-based estimates of postcensal population characteristics. *Review of Public Data Use*, 7, 16-22.
- SPSS, Inc. (1986). *SPSSx User's Guide*. Chicago: SPSS, Inc.
- STATISTIQUE CANADA (1987). *Méthodes d'estimation de la population, Canada*. N° 91-528F au catalogue, Statistique Canada.
- STOTO, M.A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78, 13-20.
- SWANSON, D. (1980). Improving accuracy in multiple regression estimates of population using principles from causal modeling. *Demography*, 17, 413-427.
- SWANSON, D., et PREVOST, R. (1986). Identifying extreme errors in ratio-correlation estimates of population. Article présenté au Annual Meeting of the Population Association of America.
- SWANSON, D., et TEDROW, L. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography*, 21, 373-381.
- TAYMAN, J., et SCHAFER, E. (1985). The impact of coefficient drift and measurement error on the accuracy of ratio-correlation population estimates. *The Review of Regional Studies*, 15, 3-10.
- U.S. BUREAU OF THE CENSUS (1977). Projections of the Population of the United States, 1977 to 2050. *Current Population Reports. Series P-25 No. 704*. Washington, D.C.: U.S. Government Printing Office.
- VERMA, R.V.P., BASAVARAJAPPA, K.G., et BENDER, R.K. (1983). Estimation par regression de la population à l'échelon infraprovincial au Canada. *Techniques d'enquête*, 9, 242-266.
- WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.

une situation où la fiabilité des estimations est à son plus faible niveau pour chaque comté. Dans ces circonstances, l'intervalle relativement large que l'on obtient pour chaque comté à un seuil de 90 % n'est pas exagéré. Le modèle pourrait subir des changements structurels au fil du temps. Ainsi, nous aurions probablement une bande de confiance qui irait en s'élargissant tout le long de la période intercensitaire qui suit l'année du dernier recensement utilisée dans la construction du modèle.

5. CONCLUSION

Il est entendu que cet article ne vise pas à présenter la méthode des rangs comme un moyen infallible de construire des intervalles de confiance pour des estimations postcensitaires de la population de petites régions. Cette méthode semble toutefois représenter une bonne base de départ pour la recherche. Malgré les faiblesses relevées, les personnes chargées d'établir des estimations postcensitaires de la population devraient tirer profit de l'utilisation du test de Wilcoxon et des intervalles de confiance construits à l'aide de la méthode des rangs. Tout d'abord, il est important, comme le soulignent Espenshade et Tayman (1983), que les utilisateurs d'estimations postcensitaires de la population aient une idée de la précision de ces estimations et c'est justement là l'objet du test de Wilcoxon et des intervalles de confiance. Deuxièmement, grâce à la construction d'intervalles de confiance acceptables, on dispose d'un moyen formel pour résoudre tout désaccord concernant la population d'une petite région donnée en ayant recours à des tests d'hypothèses. Troisièmement, l'E.T.E. peut servir de critère pour le choix d'un modèle. Cela signifie qu'il est possible d'envisager une série de modèles de corrélation des rapports pour n'importe quelle année d'estimation postcensitaire donnée et qu'en outre, on dispose d'un critère formel pour le choix du modèle. Ces éléments pourraient être utiles lorsque les estimations produites à l'aide de la méthode de corrélation des rapports par le service de la démographie d'une administration publique sont contestées pour une année postcensitaire donnée, phénomène devenu plus fréquent, particulièrement aux E.-U. (D'Allesandro 1987).

REMERCIEMENTS

L'auteur tient à remercier Jeff Tayman ainsi que des arbitres anonymes et le comité de rédaction pour leurs commentaires et suggestions. Il tient aussi à exprimer sa reconnaissance à Peggy Jobe, qui a tapé la version préliminaire de cet article, et à Carey Taylor, qui a tapé la version finale.

BIBLIOGRAPHIE

- CHOW, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28, 591-605.
- D'ALLESANDRO, F. (1987). Should applied demographers take out liability insurance? Article présenté au Annual Meeting of The Population Association of America.
- D'ALLESANDRO, F., et TAYMAN, J. (1980). Ridge regression for population estimation: Some insights and clarification. Staff Document No. 56. Office of Financial Management, State of Washington: Olympia, Washington.
- DRAPER, N.R., et SMITH, H. (1981). *Applied Regression Analysis, 2nd Edition*. New York: Wiley.
- ERICKSEN, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

Tableau 2
Nombre (%) de comtés dont la population recensée en 1970
est incluse dans l'intervalle de confiance

Niveau de confiance	E.T.E. non modifiée (0.05022)	E.T.E. modifiée (0.0599)
Intervalles moyens (en pourcentage)		
90%	35 (89.7%)	38 (97.4%)
80%	28 (71.8%)	33 (84.6%)
70%	24 (61.5%)	29 (80.6%)
66.66%	24 (61.5%)	26 (66.6%)
60%	22 (56.4%)	23 (59.0%)
50%	20 (51.3%)	22 (56.4%)
90%	9.10	10.81
80%	7.02	8.38
70%	5.66	6.75
66.66%	5.59	6.40
60%	4.59	5.47
50%	3.66	4.35

Le second aspect du tableau 2 est l'intervalle moyen qui correspond à un niveau de confiance donné. À un seuil de 90 %, l'intervalle moyen est de 9.10 % selon le modèle non modifié et de 10.81 % selon le modèle modifié. À un seuil de 50 %, il est de 3.66 % et de 4.35 % respectivement. De toute évidence, avec des niveaux de confiance de 50 et de 60 %, on obtient des intervalles moyens qui sont plus en rapport avec l'écart moyen en pourcentage, qui est de 4.88 pour le modèle modifié.

Lorsqu'on étudie la question des intervalles de confiance, on sent la nécessité d'élaborer une méthode de construction d'intervalles de confiance qui produira des intervalles fiables en ce qui a trait à la précision des estimations postcensitaires de la population de petites régions. En revanche, il faut aussi savoir choisir un niveau de confiance conforme aux estimations. Surlignons à cet égard l'étude de Stoto (1983) sur les intervalles de confiance empiriques pour des projections démographiques. Stoto (1983) a constaté notamment que les projections démographiques les plus faibles produites par le U.S. Bureau of the Census (1977) pour les États-Unis correspondaient à un intervalle de confiance à 66.66 %. Il se peut que ce niveau convienne pour des estimations postcensitaires de la population de petites régions même si, dans le test que nous avons effectué, il produit un intervalle moyen de 6.4 % selon le modèle modifié, ce qui est quelque peu supérieur à l'écart moyen (4.9 %). Il y a aussi la question du délai entre l'année pour laquelle on veut établir des estimations postcensitaires et l'année du dernier recensement. Dans l'exemple que nous venons d'étudier, on a utilisé la période postcensitaire maximum prévue aux États-Unis, soit 10 ans. Essentiellement, nous sommes devant

En utilisant les estimations calculées dans l'Etat de Washington (Swanson 1980), nous obtenons

$$R^2 = (0.07533) (0.75290) + (0.47085) (0.92146) + (0.49481) (0.88082) = 0.926$$

et

$$(S.E.E.) = \left[\frac{39.2}{(39) (0.2145)^2 (1 - 0.926)} \right]^{1/2} = 0.0599.$$

Notons que, d'après le tableau 2 de Swanson (1980), la valeur réelle du coefficient de détermination et de l'E.T.E. est 0.878 et 0.05077 respectivement. L'E.T.E. estimée est donc supérieure à l'E.T.E. réelle. Cela est normal étant donné que nous sommes moins sûrs de la précision des estimations établies par la méthode des rangs que de celle des estimations établies à l'aide du "vrai" modèle, si, bien sûr, ce modèle peut être déterminé. Par la méthode des rangs, nous pouvons maintenant construire une bande de confiance à l'aide de la formule suivante:

$$Y_i \pm (t_{37,\alpha/2}) (0.0599).$$

Le tableau 1 donne également les limites d'intervalle de confiance à 90 % pour les estimations de la population de comté de 1970. Ces intervalles de confiance sont définis par la formule:

$$\left[\frac{P_{1960}}{2522141} \right] (3032053) \left[(R_{1970}) \pm (1.69) (0.0599) \right].$$

Lorsqu'on compare les intervalles de confiance du tableau 1 à la population recensée cor-

respondante, on s'aperçoit que dans un seul cas (comté de Kittitas), la population recensée n'est pas incluse dans l'intervalle à 90 %. De fait, elle excède de 687 la limite supérieure de l'intervalle. À un niveau de confiance de 90 %, les intervalles sont relativement larges; moyenne de ± 10.81 %, minimum de ± 7.39 % (comté d'Island) et maximum de ± 12.83 % (comté de Lincoln). Si on compare ces chiffres à l'écart moyen en pourcentage pour les estimations de 1970, qui est de 4.89 (Swanson 1980), on en déduit qu'un seuil de 90 % donne des intervalles trop étendus au point de vue pratique. On aurait donc avantage à choisir un niveau de confiance plus approprié. On aurait aussi avantage à étudier les conséquences de l'utilisation de l'E.T.E. non modifiée (0.05022) tirée du modèle pour la période 1950-1960. On s'attend que le modèle non modifié produise des intervalles de confiance trop "optimistes", c'est-à-dire qu'à un niveau de confiance donné, le nombre de comtés dont la population réelle se trouve dans l'intervalle soit moins élevé que prévu. Pour approfondir ces questions, nous avons cons-

truit le tableau 2. Le tableau 2 renferme deux séries de données distinctes. Dans les deux cas toutefois, il s'agit d'une comparaison entre les estimations d'un modèle modifié et d'un modèle non modifié et entre les intervalles de confiance correspondants. En ce qui a trait à la question des intervalles de confiance optimistes pour les estimations de 1970 établies à l'aide du modèle non modifié, le tableau 2 indique que pour des niveaux de confiance allant de 50 à 90 %, les intervalles sont effectivement optimistes puisqu'il n'y a que deux niveaux de confiance pour lesquels le nombre de comtés est conforme au niveau de précision voulu. À un seuil de 80 %, par exemple, l'intervalle contient la population recensée de 28 comtés seulement (72 %); à un seuil de 60 %, l'intervalle contient la population recensée de 22 comtés seulement (56 %).

Tableau 1
Intervalles de confiance à 90% pour les estimations
de la population civile de moins de 65 ans,
par comté, État de Washington, 1970

Comté	Population recensée	Limite inférieure	Population estimée	Limite supérieure	Intervalle de confiance à 90% (en pourcentage)
Adams	11102	10335	11458	12581	± 9.80
Asotin	11862	10469	11814	13154	± 11.38
Benton	63144	60405	67511	74616	± 10.53
Chelan	35862	31733	36177	40620	± 12.28
Clallam	30023	28063	31294	34525	± 10.32
Clark	116663	101183	111437	121690	± 9.20
Columbia	3771	3683	4161	4639	± 11.49
Cowlitz	62586	55170	61581	67992	± 10.41
Douglas	15287	14569	16252	17935	± 10.36
Ferry	3336	2963	3397	3831	± 12.78
Franklin	23983	21960	24631	27302	± 10.84
Garfield	2546	2447	2761	3075	± 11.37
Grant	38921	37561	42606	47651	± 11.84
Grays Harbor	52583	46294	52114	57935	± 11.17
Island	20589	20512	22148	24040	± 7.39
Jefferson	9235	8440	9473	10506	± 10.90
King	1054271	935664	1037937	1140203	± 9.85
Kitsap	86529	77022	85821	94619	± 10.25
Kititas	22764	17649	19863	22077	± 11.15
Klickitat	10729	10440	11923	13406	± 12.44
Lewis	39265	35747	40122	44497	± 10.90
Lincoln	8168	7939	9107	10275	± 12.83
Mason	18411	16057	17827	19596	± 9.93
Okanogan	22952	21002	23795	25688	± 10.97
Pacific	13310	11270	12795	14320	± 11.92
Pend Oreille	5185	5147	5893	6639	± 12.86
Pierce	339048	314272	346728	379184	± 9.36
San Juan	3089	2636	2918	3201	± 9.66
Skagit	45703	43255	48758	54261	± 11.29
Skamania	5330	4787	5358	5929	± 10.66
Snohomish	245193	213164	231996	250827	± 8.12
Spokane	251057	227372	256723	286072	± 11.43
Stevens	15178	13869	15780	17692	± 12.11
Thurston	68719	63644	69540	75436	± 8.48
Wahkiakum	3137	3033	3397	3761	± 10.72
Walla Walla	36608	33727	38271	42812	± 11.87
Whatcom	72111	63218	70670	78122	± 10.54
Whitman	34843	28960	32409	35858	± 10.64
Yakima	128960	120347	136203	152219	± 11.69

et

$$U.L. (P^{it+x}) = \left[\frac{P^{it}}{\Sigma P^{it}} \right] (\Sigma P^{it+x}) \left[(R^{it+x}) + (t^{n-2, \alpha/2}) (S.E.E.) \right]$$

4. ETUDE EMPIRIQUE

Le tableau 1. A de Swanson (1980) donne les coefficients de corrélation totale pour un modèle de corrélation des rapports conçu pour estimer la population civile de moins de 65 ans des comtés de l'Etat de Washington pour la période 1950-1960 et ce, à partir de la liste des cotisants à l'assurance-chômage, de la liste électorale et de la liste des effectifs scolaires pour les huit premiers niveaux. Le tableau 1. B présente les caractéristiques du modèle construit à l'aide de ces données tandis que les tableaux 2. A et 2. B renferment les données correspondantes pour la période 1960-1970. Cette dernière série de données représente les données d'estimation qui serviront à illustrer la méthode.

Bien que nous connaissions toutes les données d'estimation, nous allons appliquer la méthode comme si nous ne les connaissions pas toutes. Evidemment, comme dans tout problème d'estimation, nous connaissons la matrice des coefficients de corrélation totale pour les variables indépendantes, laquelle sert, avec le théorème fondamental de l'analyse des coefficients de régression, à estimer les coefficients du modèle modifié. Si nous appliquons intégralement la méthode des rangs, nous obtenons le modèle modifié (Swanson 1980):

$$Y = 0.046618 + 0.066786X_1 + 0.50727X_2 + 0.38736X_3.$$

Le tableau 1 donne, par comté, l'estimation de la population civile de moins de 65 ans pour 1970 (redressée en fonction de l'estimation indépendante du total de la population de l'Etat), établie à l'aide du modèle modifié, de même que la population recensée.

Nous avons appliqué le test de Wilcoxon au moyen de la procédure contenue dans la commande "NPAR Tests" du SPSSX (SPSS 1986). Pour des raisons d'économie d'espace, nous ne présenterons ici ni les estimations non modifiées ni les estimations modifiées. On les trouvera dans le tableau 3 de Swanson (1980). Selon l'hypothèse nulle, la probabilité que $Z = -3.2096$ est 0.0013. Nous rejetons donc l'hypothèse nulle et supposons qu'il n'y a pas de continuité pour les comtés de l'Etat de Washington entre le modèle construit à l'aide des données de la période 1950-1960 et le vrai modèle inconnu rattaché aux données de la période 1960-1970.

Il est intéressant de noter que le test de Chow (Chow 1960) a confirmé les résultats du test de Wilcoxon en montrant que la différence entre le "vrai" modèle de corrélation des rapports (1960-1970) et le modèle pour la période 1950-1960 était statistiquement significative.

Si les résultats du test de Wilcoxon nous avaient amenés à ne pas rejeter l'hypothèse nulle, nous aurions établi les estimations de la population des comtés de l'Etat de Washington pour 1970 à l'aide des coefficients non modifiés du modèle pour la période 1950-1960. En outre, nous nous serions servis de l'E.T.E. établie pour ce modèle (0.05022) pour construire les intervalles de confiance correspondants. Or, dans ce cas précis, les résultats du test de Wilcoxon nous ont amenés à rejeter l'hypothèse nulle, ce qui indique qu'il faudrait utiliser les coefficients modifiés, déterminés à l'aide de la méthode des rangs, au lieu des coefficients non modifiés. De plus, cela vient confirmer la nécessité de définir une nouvelle E.T.E., qui n'exagèrera pas la précision des estimations pour 1970.

Pour répondre à cette question, considérons que nous examinons des paires de variables liées entre elles. Cela implique que l'on peut utiliser le test de Wilcoxon pour les observations appariées (Mosteller et Rourke 1973). L'hypothèse nulle dans ce test est qu'il n'y a aucune différence entre les estimations de la population établies à l'aide du modèle de régression normal et celles établies à l'aide du modèle de régression modifié.

Pour bien comprendre la façon dont on construit des intervalles de confiance pour des estimations postcensitaires de la population de petites régions, précisons que la méthode des rangs permet d'obtenir une série de coefficients de régression pour la série de données d'estimation. Dans un deuxième temps, on se sert de ces coefficients pour calculer la valeur estimée de R^2 et de l'E.T.E. pour la série de données d'estimation, puis on se sert finalement de l'E.T.E. estimée pour construire les intervalles de confiance. Rappelons premièrement que le coefficient de détermination multiple, R^2 , est simplement la somme des produits du coefficient de corrélation totale entre une variable dépendante et la variable dépendante par le coefficient de régression normalisé pour cette variable indépendante (Hayes 1973), de sorte que l'E.T.E. est définie (Hayes 1973)

$$S.E.E. = \left[\frac{(n) (S_y^2) (1 - R^2)}{n - 2} \right]^{1/2}$$

où

$$\begin{aligned} n &= \text{nombre de cas (petites régions)} \\ S_y^2 &= \text{variance de la variable dépendante} \\ R^2 &= \text{coefficient de détermination multiple.} \end{aligned}$$

Selon Kmenta (1971), la formule de l'intervalle de confiance pour une valeur estimée donnée sur une droite de régression (population) est

$$Y_i \pm (t_{n-2,\alpha/2}) (S.E.E.).$$

Il faut bien se rendre compte que l'intervalle de confiance obtenu ne porte pas sur une estimation de la population mais bien sur une estimation d'un rapport de proportions ou $R_{it} + x$. Or, Espenshade et Tayman (1982) ont montré qu'un intervalle de confiance construit pour une variable pouvait être adapté à une autre variable substituée algébriquement à la première. Par conséquent, en déterminant les limites de confiance de $R_{it} + x$, nous pouvons les exprimer sous la forme de valeurs de population:

$$\begin{aligned} &(R_{it+x}) \pm (t_{n-2,\alpha/2}) (S.E.E.) \\ &= \left[\frac{P_{it+x}}{P_{it}} \right] \div \left[\frac{P_{it}}{\Sigma P_{it}} \right] \pm (t_{n-2,\alpha/2}) (S.E.E.) \\ &L.L. (P_{it+x}) = \left[\frac{P_{it}}{\Sigma P_{it}} \right] (P_{it+x}) \left[(R_{it+x}) - (t_{n-2,\alpha/2}) (S.E.E.) \right] \end{aligned}$$

ce qui donne

i = petite région ($1 \leq i \leq n$)

t = année du dernier recensement

et

(1.A)
$$R_{it} = \left[\frac{\sum P_{it}}{\sum P_{it-z}} \right] \div \left[\frac{\sum S_{it}}{\sum S_{it-z}} \right]$$

(1.B)
$$(X_i)_{itj} = \left[\frac{\sum S_{it}}{\sum S_{it}} \right] \div \left[\frac{\sum S_{it-z}}{\sum S_{it-z}} \right]_j$$

ou

Z = nombre d'années entre les recensements

P = population

S = indicateur symptomatique.

Une fois le modèle construit, on établit une estimation postcensitaire pour la période $t + x$ en substituant $(S_{i,t+x}/\sum S_{i,t+x})_j$ au numérateur du membre de droite de l'équation [1.B] et $(S_{it}/\sum S_{it})_j$ au dénominateur. Ainsi, une fois que l'on a déterminé $R_{i,t+x}$, on calcule la population réelle de la région i à la période $t + x$ en introduisant dans l'équation, [1.A] une estimation indépendante du total de la population, P_{t+x} , puis en solutionnant algébriquement cette équation en fonction de $P_{i,t+x}$. Comme $\sum P_{i,t+x}$ égale rarement l'estimation indépendante du total de la population P_{t+x} , on effectue une correction de manière à ce que cette estimation corresponde à la somme des chiffres de population.

Une faiblesse de la méthode de corrélation des rapports est l'invariabilité de sa structure. C'est ce qui a amené Swanson (1980) à définir la méthode des rangs. Cette méthode repose sur l'idée, mise de l'avant par Land (1969, Chapitre IV), que l'on peut exploiter l'information contenue dans les coefficients de corrélation totale qui se trouvent dans une série de données d'estimation. L'étude de Land repose sur le théorème fondamental de l'analyse des coefficients de direction énoncé par Wright (1921). Dans cette étude, il est question d'une transformation théorique de la variable dépendante du modèle de régression (variable de population) en une variable explicative non mesurée et d'une structure jusqu'à l'implicité – minimum de trois variables explicatives (dans le modèle de régression), dont la covariance peut être rattachée au fait qu'elles sont elles-mêmes expliquées par la variable de population.

3. MÉTHODE D'ESTIMATION DES INTERVALLES DE CONFIANCE

Si la stabilité des liens observés entre les variables dans la série de données du modèle est vérifiée (comme permet de le faire la méthode des rangs), on devrait retrouver les mêmes liens entre les variables dans la série de données d'estimation. Cette stabilité indique que l'on peut se fonder sur l'erreur type d'estimation ($E.T.E.$) rattachée à la série de données du modèle pour construire des intervalles de confiance pour la série de données d'estimation. En revanche, si la stabilité n'est pas vérifiée, l' $E.T.E.$ rattachée à la série de données du modèle ne convient pas et peut, de fait, produire des intervalles de confiance qui exagèrent la précision des estimations postcensitaires. Ces considérations nous amènent à nous demander s'il n'y aurait pas lieu de vérifier la stabilité des liens entre les variables par l'inférence statistique.

La méthode des rangs, version modifiée de la méthode de corrélation des rapports définie par Swanson (1980), qui applique les principes de la modélisation causale pour tenir compte des changements structurels postcensitaires dans un modèle de corrélation des rapports donné.

La construction d'intervalles de confiance par la méthode de corrélation des rapports soulève trois questions fondamentales. La première concerne la stabilité temporelle du modèle. Si la structure des liens qui existent entre les variables du modèle est fixe dans le temps, les intervalles de confiance construits par rapport à la série de données du modèle s'appliquent-ils aux estimations de la population produites par le modèle à partir de la série de données d'estimation. Même si de nombreux auteurs s'entendent pour dire qu'il n'est pas prudent de supposer l'invariance d'un modèle (D'Allesandro et Tayman 1980; Ericksen 1973, 1974; Mandell et Tayman 1982; Namboodiri 1972; O'Hare 1976, 1980; Smith et Mandell 1984; Spar et Martin 1979; Swanson 1980; Swanson et Prevost 1986; Swanson et Tedrow 1984; Tayman et Schaffer 1982; Verma et coll. 1983), il serait utile de disposer d'un test de stabilité. Cela nous amène à la seconde question, soit l'utilisation d'un test statistique. Si le test indique que l'on ne peut supposer la stabilité du modèle et que, malgré cela, on applique, par exemple, des intervalles de confiance fondés sur un modèle construit à l'aide de données de la période 1960-1970 à des estimations produites pour 1979, ces intervalles sont susceptibles d'exagérer le niveau de précision des estimations de 1979. D'où la nécessité de trouver une méthode qui produira des intervalles de confiance acceptables. Ce dernier point représente la troisième question relative à la construction d'intervalles de confiance par la méthode de corrélation des rapports.

Dans les sections qui suivent, nous décrivons la méthode de corrélation des rapports ainsi que sa version modifiée à partir de laquelle nous pourrions construire des intervalles de confiance acceptables. Nous décrivons également toute la logique qui sous-tend la construction de ces intervalles de confiance, puis nous passons à un exemple empirique qui illustre le test de stabilité et la production d'intervalles de confiance "satisfaisants" et "insatisfaisants".

2. MÉTHODE D'ESTIMATION DE LA POPULATION

La méthode de corrélation des rapports est une méthode de régression qui sert à mesurer la variation temporelle des proportions de la population de petites régions à partir de la variation observée dans les proportions d'indicateurs symptomatiques comme ceux tirés de la liste électorale, de la liste des cotisants à l'assurance-chômage ou de la liste des effectifs scolaires. On mesure la variation temporelle en calculant simplement un rapport de proportions pour deux périodes différentes. Comme les chiffres de population pour toutes les petites régions ne peuvent être connus que par le recensement fédéral, on construit toujours un modèle de régression de la corrélation des rapports en utilisant deux périodes séparées par un intervalle régulier. Le modèle est défini par la formule

$$R_{it} = a_0 + \sum_{j=1}^k (b_j) (X_{jt})_i + \epsilon$$

où

a_0 = ordonnée à l'origine à estimer

b_j = coefficient de régression à estimer

ϵ = terme d'erreur

j = indicateur symptomatique, ($1 \leq j \leq k$)

Intervalle de confiance pour les estimations postcensitaires de la population: une étude de cas pour les petites régions

DAVID A. SWANSON¹

RÉSUMÉ

L'auteur présente une méthode qui permet de construire des intervalles de confiance acceptables pour des estimations postcensitaires de la population en utilisant une version modifiée de la méthode de corrélation des rapports, appelée méthode des rangs. Il montre que le test de Wilcoxon peut servir à déterminer si un modèle de corrélation des rapports donné est stable à long terme. Si la stabilité du modèle est vérifiée, on en conclut que les intervalles de confiance liés aux données qui ont servi à la construction du modèle sont acceptables pour des estimations postcensitaires. Dans le cas contraire, on conclut que les intervalles de confiance ne sont pas acceptables et qu'en plus, ils sont susceptibles d'exagérer la précision des estimations postcensitaires. Étant donné un modèle instable, l'auteur montre que l'on peut néanmoins déterminer des intervalles de confiance acceptables pour les estimations postcensitaires en utilisant la méthode des rangs. Il présente finalement un exemple empirique où il utilise les estimations de la population des comtés de l'État de Washington.

MOTS CLÉS: Estimation de la population; intervalles de confiance; méthode de régression de la corrélation des rapports.

1. INTRODUCTION

Il n'existait pas de méthode de construction d'intervalles de confiance pour des estimations postcensitaires avant que Espenshade et Tayman (1982) n'élaborent une méthode d'estimation par âge. L'introduction de cette méthode représente une étape majeure dans l'évolution des techniques d'estimation; néanmoins, comme la plupart des techniques de pointe, elle présente des lacunes, dont deux méritent d'être soulignées:

1. la méthode risque de ne pas donner de résultats satisfaisants au niveau infra-provincial ou infra-étatique (Espenshade et Tayman 1982);

2. elle s'éloigne beaucoup de la méthode de régression normalement utilisée au Canada et aux États-Unis pour estimer la population de petites régions, c'est-à-dire de la méthode de corrélation des rapports; ce point est très important car il soulève la question de la nature des données qu'il faudrait alors recueillir et de la nouveauté de cette méthode pour les personnes chargées d'estimer la population des petites régions (Statistique Canada 1987). Le terme "petite région" désigne une division de recensement au Canada (Statistique Canada 1987). Étant l'Alaska, où les petites régions sont définies comme des secteurs de recensement, la Louisiane, où les paroisses sont l'équivalent des comtés, et la Virginie, où les "villes autonomes" sont considérées comme des petites régions.

Dans cet article, nous allons exposer une méthode qui permet de construire des intervalles de confiance pour des estimations postcensitaires de la population de petites régions en utilisant

¹ David A. Swanson, Département de sociologie, Pacific Lutheran University, Tacoma, Washington 98447, U.S.A.

données sur la migration sont comparables et produisent des estimations provinciales similaires, d'erreurs en fin de période de mêmes niveaux de variation d'une province à l'autre. Puisque les coefficients de variation sont inférieurs à 20%, les deux types de fichiers fournissent des données acceptables sur la migration interne.

En conclusion, les estimations de la migration interprovinciale issues des fichiers d'impôt de Revenu Canada et des allocations familiales sont cohérentes pour la période 1981-86. Au niveau provincial, on observe une plus faible variation dans les erreurs en fin de période.

4. CONCLUSION ET DISCUSSION

Les fichiers d'allocations familiales et d'impôt de Revenu Canada jouent un rôle important dans l'établissement d'estimations cohérentes de l'émigration et de la migration interne pour le Canada, les provinces et les territoires. Les estimations des émigrants et des migrants interprovinciaux produites à partir de ces fichiers pour la période 1981-1986 sont acceptables pour fins d'estimation de la population totale.

Pourtant, un problème demeure. Au niveau national, l'erreur en fin de période (l'écart entre les estimations de population et les effectifs recensés) pour 1986 était supérieure à celle relevée pour les trois recensements antérieurs, ceux de 1971, 1976 et 1981. De plus, toujours en 1986, pour toutes les provinces, les estimations de la population étaient supérieures aux chiffres du recensement.

Ces irrégularités sont, dans une large part, le résultat des différences de complétude des recensements de 1981, qui a servi de population repère, et de 1986. La contre-vérification des dossiers estimait, en 1981, le taux de sous-dénombrement pour le Canada à 2,01%. En 1986, l'estimation était beaucoup plus forte, 3,21%.

Les erreurs dans les estimations des autres composantes de l'accroissement démographique peuvent aussi, du moins en partie, être responsables de ces irrégularités.

REMERCIEMENTS

Nous tenons à remercier les évaluateurs de cet article pour leurs commentaires et suggestions.

BIBLIOGRAPHIE

BEAUJOT, R., et RAPPAPAK, J.P. (1988). *L'émigration du Canada: importance et interprétation*. Ottawa: Emploi et Immigration Canada.

NORRIS, D., BRITTON, M., et VERMA, R.B.P. (1982). The use of administrative records for estimating migration and population. *Statistics of Income and Related Administrative Record Research*: 1982, Washington, D.C.: Department of the Treasury, Internal Revenue Service.

NORRIS, D., et STANDISH, L.D. (1983). *Rapport technique sur la production de données migratoires à partir des dossiers d'impôt*. Rapport technique. Ottawa: Division des données régionales et administratives, Statistique Canada.

RABY, R., MARTEL, J., et CARTIER, G. (1989). *Questions relatives aux estimations postcensitaires courantes*. Document présenté au comité fédéral-provincial de démographie, Ottawa.

STATISTIQUE CANADA (1987). *Méthodes d'estimation de la population, Canada*. No 91-528F au catalogue.

VERMA, R.B.P., et PARENT, P. (1985). Vue d'ensemble des avantages et inconvénients des fichiers de données administratives choisis. *Techniques d'enquête*, 11, 193-202.

Erreur en fin de période entre les divers genres d'estimations de la population et les chiffres du recensement, par province et le territoire, 1971, 1976, 1981 et 1986

Tableau 4

Région géographique	Erreur en fin de période ¹ (%)			
	1971	1976	1981	1986
	Impôt	A.F.	Impôt	A.F.
	Impôt	Impôt	Impôt	Impôt

Terre-Neuve	-2.08	-1.64	0.49	1.34	1.63	2.30	1.97	2.01
Ile-du-Prince-Edouard	-2.09	-2.01	0.17	2.11	-0.05	1.02	0.99	0.63
Nouvelle-Ecosse	-1.68	-2.39	-0.20	1.18	0.30	0.40	1.24	1.04
Nouveau-Brunswick	-1.93	-2.65	-1.29	1.81	0.13	0.54	1.58	1.04
Québec	-0.33	-0.97	-0.05	-0.18	-0.30	-0.07	1.32	1.40
Ontario	0.11	0.99	0.15	0.16	0.64	0.37	0.72	0.65
Manitoba	0.29	0.38	-0.27	0.39	1.07	0.87	0.51	0.41
Saskatchewan	0.44	-0.33	0.45	0.37	-0.31	0.28	1.08	1.31
Alberta	-0.14	0.52	-1.07	-1.11	-2.39	-2.64	0.73	0.63
Colombie-Britannique	0.01	-1.34	0.28	-1.10	0.03	-0.07	0.59	0.79
Yukon	-5.36	-5.99	-0.87	3.79	-1.98	2.06	-4.78	-3.10
Territoires du Nord-Ouest	-2.12	2.64	-12.98	-3.39	-7.08	0.43	-1.44	-1.40
Provinces et territoires	0.91	1.33	0.44	0.97	0.69	0.86	1.07	1.01
10 provinces	1.38	1.82	1.52	1.41	1.33	0.92	1.41	1.22

Nota: De 1976 à 1980, les données de Revenu Canada pour les enfants n'étaient disponibles que pour le groupe d'âge de 0 à 15 ans. Par conséquent, les facteurs $f_{(j,k)}$ ont été calculés au moyen du nombre de migrants âgés de 0 à 15 ans et de 16 ans et plus et non de 0 à 17 ans et de 18 ans et plus.

¹ L'erreur en fin de période est calculée au moyen de l'équation suivante:

$$\text{Erreur en fin de période} = \left(\frac{\text{Estimation} - \text{Recensement}}{\text{Recensement}} \right) \times 100$$

Sources: Impôt: Fichier d'impôt de Revenu Canada. A.F.: Fichier des allocations familiales. Sources: Estimations de la migration interprovinciale basées sur le fichier des allocations familiales, Division de la démographie, Statistique Canada. Estimations de la migration interprovinciale basées sur des données d'impôt, Division des données régionales et administratives, Statistique Canada.

Tableau 5

Coefficients de variation de l'erreur absolue moyenne en fin de période entre les estimations de la population et les chiffres du recensement pour les provinces ($n = 10$), selon la source des estimations de la migration interprovinciale, 1966-1971, 1971-1976, 1976-1981 et 1981-1986

Période	Source	EAM	(t + 5)	Ecart-type	Coefficient de variation (%)
---------	--------	-----	---------	------------	------------------------------

1966-1971	Impôt	0.91	(1)	(2)	(3) = (2 ÷ 1) × 100
	A.F.	0.2863			31
1971-1976	Impôt	1.33			20
	A.F.	0.44			30
1976-1981	Impôt	0.97			22
	A.F.	0.2135			36
1981-1986	Impôt	0.69			33
	A.F.	0.2855			14
	Impôt	1.07			16
	A.F.	0.1570			

Nota: EAM: Erreur absolue moyenne en fin de période.
Impôt: Fichier d'impôt de Revenu Canada.
A.F.: Fichier des allocations familiales.
Source: Division de la démographie, Statistique Canada.

Tableau 3

Estimations de la migration interprovinciale nette d'après la question de mobilité du recensement de 1986, le fichier des allocations familiales, le fichier d'impôt et la méthode résiduelle, Canada, provinces et territoires, 1981-1986

Région géographique	Recensement de 1986 ¹	Allocations familiales	Impôt sur le revenu	Méthode résiduelle ²
CANADA	0	0	0	-238,178
Terre-Neuve	-16,550	-14,837	-15,051	-26,111
Ile-du-Prince-Edouard	1,540	293	751	-509
Nouvelle-Ecosse	6,275	5,204	6,895	-4,095
Nouveau-Brunswick	-1,370	-2,239	-65	-11,212
Québec	-63,295	-76,040	-81,254	-167,286
Ontario	99,355	115,497	121,767	57,147
Manitoba	-1,555	-3,700	-2,634	-8,180
Saskatchewan	-2,820	-668	-2,974	-13,564
Alberta	-27,665	-34,073	-31,676	-50,811
Colombie-Britannique	9,500	13,289	7,382	-12,418
Yukon	-2,665	-2,381	-2,775	-1,643
Territoires du Nord-Ouest	-755	-345	-366	504

¹ Population âgée de 5 ans et plus.
² La méthode résiduelle d'estimation de la migration interprovinciale nette est la suivante:
Migration nette = accroissement de la population recensée entre t et $t + 5$ - [(naissances - décès) + (immigration - émigration)].
Source: Division de la démographie, Statistique Canada.

estimations des autres composantes de l'accroissement démographique sont exactes. Sur le plan statistique, un coefficient de variation se situant entre 20% et 30% est généralement considéré comme acceptable.

Cependant, certains pourraient nous opposer que le coefficient de variation n'est pas un bon indicateur de la qualité des données de migration interne. Par exemple, une série d'estimations dont l'erreur absolue en fin de période est de 10% pour chaque province générerait un coefficient de variation nul et en conséquence, serait préférable à une série d'estimations d'erreurs en fin de période étalées entre -1,0% et 1,0%. Dans de tels cas, une mesure qualitative qui prend en compte tant la taille de l'erreur absolue en fin de période que leur écart-type est clairement nécessaire. Cependant, la probabilité que les provinces aient la même erreur absolue en fin de période est extrêmement faible (voir le tableau 5), donc, l'application du coefficient de variation dans ce texte nous semble valable.

Le tableau 5 montre les coefficients de variation (calculés d'après les chiffres du tableau 4) entre d'une part les estimations de la population fondées sur les deux séries d'estimations de migration interne et d'autre part les chiffres des recensements de 1971, 1976, 1981 et 1986. Avant 1976, les coefficients de variation des données de migration tirées du fichier d'impôt étaient supérieurs de 50% à ceux correspondant au fichier d'allocations familiales. Il fallait s'attendre à cette observation puisque la méthode d'estimation de la migration à partir du fichier d'impôt n'était, à ce moment-là, qu'à l'étape d'élaboration. De plus, le facteur f_j (rapport du nombre d'adultes émigrants à celui des enfants) servant à l'estimation du nombre de migrants interprovinciaux était basé sur les données de mobilité du recensement, approche jugée depuis lors moins satisfaisante que la méthode actuelle. C'est pourquoi on constate, pour les périodes 1976-81 et 1981-86, une nette diminution des écarts entre les coefficients de variation correspondant aux deux types de fichiers.

En 1981, le coefficient de variation correspondant au fichier d'impôt était supérieur de 9% à celui relatif au fichier des allocations familiales, tandis qu'en 1986, il lui était inférieur de 12%. Comme ces différences sont minimales, nous pouvons affirmer que les deux séries de

Puisque, selon certains auteurs (Beaufort et Rappak 1988), il y a corrélation entre les flux d'émigrants et d'immigrants, on peut calculer le facteur f_c au moyen du fichier d'Emploi et Immigration Canada (EIC). Les valeurs du facteur f_c établies au moyen du fichier d'immigration d'EIC se situent entre les valeurs basées sur les données de migration interprovinciale et celles correspondant aux émigrants vers les États-Unis. L'effectif basé sur le facteur f_c déduit du fichier d'immigration (275,762) est supérieur à l'estimation officielle des émigrants (235,481), mais est voisin de celle obtenue à partir de l'étude de contre-vérification des dossiers de 1986 (288,376). Si l'estimation officielle du nombre d'émigrants était accrue à 275,762, l'erreur en fin de période en 1986 entre l'estimation de la population et l'effectif recensé serait réduite de 0,95% à 0,79%.

En résumé, pour la période 1981-86, il semblerait qu'on puisse améliorer les estimations de l'émigration en basant le calcul du facteur f_c non plus sur le fichier d'impôt de Revenu Canada mais plutôt sur les données d'Emploi et Immigration Canada. Déjà en mars 1989, il est apparu que les estimations d'émigrants basées sur les fichiers des allocations familiales et un facteur f_c calculé à partir des données d'immigration d'EIC demeuraient trop faibles après 1986. Ceci semble résulter de la forte proportion (33%) d'émigrants canadiens vers les États-Unis de 1981 à 1986, selon les données américaines.

Une analyse a également été faite d'une méthode combinant les effectifs d'émigrants canadiens aux États-Unis selon le U.S. Department of Justice, Immigration and Naturalization Service, les effectifs d'enfants émigrants (âgés de 0-17 ans) selon les fichiers des allocations familiales et un facteur f_c basé sur le fichier d'immigration d'EIC pour les pays autres que les États-Unis. Le nombre estimé d'émigrants de 1981 à 1986 selon cette méthode était de 285,413. Cette estimation révisée est très près de celle basée sur la contre-vérification des dossiers (288,376).

3.2 Données de migration interprovinciale

Pour vérifier la précision des estimations de la migration interprovinciale obtenues du fichier d'impôt de Revenu Canada, deux évaluations ont été menées: i) une comparaison des séries de données de la migration interprovinciale, basées sur le fichier d'impôt de Revenu Canada et sur le fichier des allocations familiales; ii) une comparaison des erreurs en fin de période des estimations de population utilisant ces deux séries de données de migration interne. Le tableau 3 compare les estimations de la migration interprovinciale nette basées sur quatre sources: la question de mobilité du recensement de 1986; le fichier d'impôt de Revenu Canada; le fichier des allocations familiales; et la méthode résiduelle. Pour chacune des provinces, les estimations de migration interne produites au moyen des données de mobilité du recensement de 1986, du fichier d'impôt de Revenu Canada et du fichier d'allocations familiales sont cohérentes, la migration nette variant toujours dans le même sens. Toutes les sources à l'exception de la méthode résiduelle montrent une migration nette positive pour l'Île-du-Prince-Édouard, la Nouvelle-Écosse, l'Ontario et la Colombie-Britannique. La migration nette des autres provinces était négative.

Les estimations de la migration interprovinciale nette calculées au moyen des fichiers d'allocations familiales et d'impôt de Revenu Canada ne sont pas strictement comparables à celles obtenues de façon résiduelle. Par définition, la somme de la migration interprovinciale nette du Canada devrait être égale à zéro. Toutefois, cette somme s'élève à 238,178 lorsqu'on utilise la méthode résiduelle. De plus, la différence entre la migration interprovinciale nette estimée de façon résiduelle et les estimations produites au moyen des fichiers d'impôt de Revenu Canada et des allocations familiales est considérable pour Terre-Neuve, le Nouveau-Brunswick, le Québec, l'Ontario et l'Alberta. On a utilisé le coefficient de variation (le rapport entre l'erreur-type de l'erreur absolue moyenne en fin de période pour les provinces et l'erreur absolue moyenne en fin de période) pour mesurer la précision relative des estimations de migration interne, en supposant que les

Tableau 2
Estimations des émigrants selon la méthode des allocations familiales, avec f_c (rapport du nombre d'adultes émigrants au nombre d'enfants émigrants) basé sur différentes sources de données, 1981-1986

Source de données de f_c	Valeur du facteur f_c				Effectifs d'émigrants
	1981-82	1982-83	1983-84	1984-85	1985-86
1. Fichier d'impôt de Revenu Canada	0.8698	0.8768	0.9052	0.8592	235,481
2. Données de migration interprovinciale tirées du fichier d'impôt de Revenu Canada	1.0760	1.1000	1.0664	1.0290	1.0029
3. Données d'immigration d'EIC	1.0801	1.0926	1.1723	1.1254	1.0694
4. Émigrants canadiens aux États-Unis	1.2300	1.2774	1.3196	1.3745	1.4232
					316,268

Source: Division de la démographie, Statistique Canada.

Pour 1976-81, les méthodes d'estimation ne produisent pas les mêmes résultats. Les effectifs estimés par la méthode résiduelle avec ajustement pour le sous-dénombrement net étaient de 194,155, valeur voisine de l'estimation basée sur le fichier d'impôt de Revenu Canada (207,420), mais beaucoup plus faible que l'estimation produite par la méthode des allocations familiales (278,624) ou par la contre-vérification des dossiers (296,724).

Une source possible d'erreur dans la méthode des allocations familiales est liée au facteur f_c , c'est-à-dire le rapport du taux d'émigration des adultes à celui des enfants, qui permet d'estimer le nombre d'émigrants de 18 ans et plus en 1981-1986. Ces facteurs sont obtenus à partir des données d'émigration fournies par le fichier d'impôt de Revenu Canada. Le tableau 2 montre les valeurs f_c selon diverses sources de données. Les facteurs f_c dérivés du fichier d'impôt de Revenu Canada sont inférieurs à 1, tandis que ceux calculés d'après les trois autres sources de données (c'est-à-dire les données sur la migration interprovinciale tirées du fichier d'impôt sur le revenu, les fichiers d'immigration et les données sur les Canadiens émigrant aux États-Unis) sont supérieurs à 1. En conséquence, le nombre d'émigrants estimé selon ces dernières sources de données est supérieur à celui basé sur le fichier d'impôt.

Pour chaque source utilisée, la valeur du facteur f_c varie d'une année à l'autre. Les valeurs du facteur f_c pour les Canadiens émigrant aux États-Unis sont relativement élevées: le nombre d'adultes émigrant vers ce pays est de 23% à 42% supérieur au nombre d'enfants émigrants. Cette observation n'est pas surprenante, puisque les États du sud des États-Unis ont toujours attiré les retraités canadiens. Par conséquent, la valeur du facteur f_c établie d'après les données sur l'émigration vers les États-Unis pourrait ne pas convenir à l'estimation des émigrants quittant le Canada vers des pays autres que le pays voisin. Les valeurs des facteurs f_c calculés à l'aide des données de migration interprovinciale obtenues du fichier d'impôt de Revenu Canada, laissent quant à elles supposer que le nombre d'adultes migrants a dépassé jusqu'à 10% le nombre d'enfants migrants entre 1981 et 1986. Cependant, chez ces adultes migrants, il pourrait y avoir eu une plus forte proportion de jeunes adultes qui ont tendance à migrer plus souvent d'une province à l'autre que les autres groupes d'âge. Par conséquent, cette source de données est aussi très spécifique et ne convient pas au calcul du facteur f_c global.

3. ÉVALUATION DES ESTIMATIONS DES COMPOSANTES DE L'ACCROISSEMENT DÉMOGRAPHIQUE

Chaque des composantes de l'accroissement démographique (les naissances, les décès, les émigrants, les émigrants interprovinciaux) est susceptible de présenter certaines erreurs. Toutefois, on peut considérer que les données relatives aux naissances, aux décès et à l'immigration sont plus précises que les estimations d'émigrants et de migrants interprovinciaux. En 1982, les méthodes d'estimation des émigrants et de la migration interne ont été remises à jour (voir Statistique Canada 1987). Ces méthodes révisées sont évaluées ci-dessous.

3.1 Données d'émigration

Le tableau 1 présente les estimations des émigrants du Canada établies selon diverses méthodes et à partir de différentes sources de données, pour les périodes 1976-1981 et 1981-1986. Pour la période 1981-1986, les effectifs des émigrants par la méthode résiduelle sont de beaucoup supérieurs à ceux dérivés de la méthode des allocations familiales. La méthode résiduelle sous-trait de l'accroissement naturel et de l'immigration la croissance de la population entre 1981 et 1986, non ajustée du sous-dénombrement au recensement. Comme les données sur les naissances, les décès et l'immigration sont considérées comme des renseignements précis, la plus forte estimation de l'émigration par la méthode résiduelle peut être attribuée à la différence dans les taux de sous-dénombrement des recensements de 1981 et 1986. Après correction des effectifs, de 2,01% pour le recensement de 1981 et 3,21% pour celui de 1986, le nombre d'émigrants estimé de façon résiduelle s'établit à 134,857. Ce nombre est le plus faible effectif estimé d'émigrants obtenu pour l'ensemble des méthodes (235,481 selon la méthode des allocations familiales et 165,272 selon les estimations basées sur le fichier d'impôt de Revenu Canada). Cette faible estimation peut résulter de taux de surdénombrement différents aux recensements de 1981 et 1986. Aucune estimation du surdénombrement n'est calculée dans l'étude de la contre-vérification des dossiers, mais on peut supposer que le taux est similaire à celui observé aux États-Unis en 1980 qui s'établit à 25% du taux de sous-dénombrement. Après correction des chiffres des recensements de 1981 et de 1986 pour les taux de couverture nette, correction de 1,51% et de 2,40% respectivement, l'estimation résiduelle des émigrants est voisine de celle dérivée de la méthode des allocations familiales, soit 218,148 relativement à 235,481.

Tableau 1

Estimations des émigrants selon différentes méthodes, Canada, 1976-1981 et 1981-1986

Méthode	1976-81	1981-86
Résiduelle*		
(a) sans ajustement	277,558	476,373
(b) avec ajustement pour le sous-dénombrement	196,955 ¹	134,857 ¹
(c) avec ajustement pour le sous-dénombrement net	194,155 ²	218,148 ²
Fichier d'impôt de Revenu Canada	207,420	165,272
Méthode des allocations familiales	278,624	235,481
Contre-vérification des dossiers	296,724	288,376

*Méthode résiduelle:
Émigrants = (naissances - décès) + [immigrants] - accroissement intercentenaire de la population entre t et t + 5.

1 Les taux de sous-dénombrement étaient de 2,04% pour le recensement de 1976, 2,01% pour celui de 1981 et 3,21% pour celui de 1986.
2 En présupposant que le surdénombrement représente 25% du taux de sous-dénombrement, comme le montre l'expérience américaine, les taux de sous-dénombrement net des recensements de 1976, 1981 et 1986 s'établissent à 1,53%, 1,51% et 2,40% respectivement.

2.3 Emigration

L'utilisation des fichiers administratifs est essentielle dans l'estimation de l'émigration, le Canada ne disposant d'aucun système de collecte de données sur les émigrants. Le fichier d'impôt de Revenu Canada permet de retracer les émigrants puisqu'ils y sont définis par une adresse "hors Canada" sur la déclaration d'une année donnée et une adresse "au Canada" pour l'année précédente. Le fichier des allocations familiales, quant à lui, permet l'identification des enfants émigrants par le biais des changements d'adresse des bénéficiaires. Les deux fichiers administratifs sont donc utilisés conjointement pour estimer les effectifs provisoires et définitifs d'émigrants. La méthode d'estimation (identifiée comme la méthode des allocations familiales) est semblable à celle utilisée pour l'estimation provisoire de la migration internationale et est définie comme suit:

$$E_j = \left[\frac{E_{j,0-17}}{P_{j,0-17}} \cdot f_c \cdot P_{j,18+} \right] + E_{j,0-17} \tag{5}$$

$$f_c = \frac{E'_{c,18+}}{E'_{c,0-17}} \div \frac{P_{c,18+}}{P_{c,0-17}} \tag{6}$$

$$E_c = \sum_{j=1}^{12} \left[E_j \right] \tag{7}$$

où:

- E_j = nombre annuel estimé d'émigrants de la province j
- E_c = nombre annuel estimé d'émigrants du Canada
- $E_{j,0-17}$ = nombre d'émigrants de la province j âgés de 0 à 17 ans inclusivement et admissibles à l'allocation familiale
- $P_{j,0-17}$ = nombre d'enfants admissibles à l'allocation familiale dans la province j
- $P_{j,18+}$ = population adulte estimée de la province j , obtenue en soustrayant le nombre d'enfants admissibles à l'allocation familiale de la population totale estimée
- f_c = facteur d'ajustement annuel, servant à estimer l'émigration des adultes selon le fichier d'impôt de Revenu Canada.
- $E'_{c,18+}$ et $E'_{c,0-17}$ = nombres estimés d'adultes et d'enfants émigrants du Canada, basés sur les fichiers d'impôt de Revenu Canada.
- $P_{c,18+}$ et $P_{c,0-17}$ = estimation de la population des adultes et des enfants pour le Canada, au 1er juin selon la méthode des composantes.

La méthode d'estimation du nombre d'émigrants a été modifiée en mars 1989, affectant les estimations postérieures à 1986. La nouvelle méthode combine les effectifs par âge des émigrants du Canada vers les États-Unis (selon le U.S. Department of Justice, Immigration and Naturalization Service) aux estimations du nombre d'émigrants du Canada vers des pays autres que les États-Unis, basées sur les fichiers des allocations familiales et un facteur f_c calculé au moyen des fichiers d'immigration (voir Raby, Martel et Cartier 1989).

(2)
$$M_{(j,k),18+} = \frac{M_{(j,k),0-17} \cdot P_{j,0-17}}{P_{j,18+} \cdot f_{(j,k)}}$$

(3)
$$f_{(j,k)} = \frac{M'_{(j,k),18+}}{M_{(j,k),0-17}} \div \frac{P_{j,18+}}{P_{j,0-17}}$$

(4)
$$M_{(j,k),0+} = M_{(j,k),18+} + M_{(j,k),0-17}$$

où :

$M_{(j,k),0+}$ = nombre total estimé de personnes quittant la province j à destination de la province k

$M_{(j,k),18+}$ = nombre estimé d'adultes (personnes âgées de 18 ans et plus) quittant la province j à destination de la province k

$M'_{(j,k),18+}$ = nombre d'adultes quittant la province j à destination de la province k selon le fichier d'impôt de Revenu Canada

$M_{(j,k),0-17}$ = nombre d'enfants (personnes âgées de 0 à 17 ans) quittant la province j à destination de la province k selon le fichier d'impôt de Revenu Canada

$M_{(j,k),0-17}$ = nombre d'enfants quittant la province j à destination de la province k , selon le fichier des allocations familiales

$P_{j,18+}$ = nombre estimé d'adultes dans la province j , différence entre l'estimation de la population totale et l'estimation de la population d'enfants selon le fichier des allocations familiales

$P_{j,0-17}$ = nombre total d'enfants recevant une allocation familiale dans la province j

$f_{(j,k)}$ = facteur d'estimation des adultes quittant la province j à destination de la province k , selon les estimations de la migration établies à partir du fichier d'impôt de Revenu Canada

$P_{j,18+}$ = nombre d'enfants dans la province j , estimations de la Division de la démographie

$P_{j,0-17}$ = nombre d'adultes dans la province j , estimations de la Division de la démographie.

2.2.2 Estimations définitives

Les estimations définitives de la migration interprovinciale sont produites différemment, et se basent sur le seul fichier d'impôt de Revenu Canada. Toutes les personnes qui gagnent un revenu annuel supérieur à une somme minimale fixée doivent remplir une déclaration de revenu aux fins d'impôt avant la fin du mois d'avril de chaque année. On peut isoler les contribuables migrants en comparant pour chaque déclarant l'adresse du domicile consignée dans les déclarations de deux années consécutives. Le nombre et l'âge des personnes à charge sont déduits du montant de l'exemption personnelle totale du contribuable. Un ajustement est ensuite fait afin de prendre en compte les personnes non couvertes par le système de Revenu Canada; ceci inclut les migrants qui ne remplissent pas de déclaration de revenu aux fins d'impôt et qui ne figurent pas parmi les dépendants sur la déclaration d'un autre contribuable (Norris et Standish 1983; Statistique Canada 1987).

2. SOURCES DE DONNÉES ET MÉTHODES D'ESTIMATION

Cette section décrit les procédures d'estimation de la population totale, de la migration inter-provinciale et de l'émigration.

2.1 Population totale

Les estimations trimestrielles et annuelles de la population totale du Canada, des provinces et territoires, de même que les totaux annuels des divisions et régions métropolitaines de recensement sont produits par la méthode des composantes. À l'échelon national, le nombre de naissances et d'immigrants est ajouté à, et le nombre de décès et émigrants soustrait à la population de base (l'effectif du recensement canadien le plus récent). Pour les provinces et les régions plus petites, les estimations de la migration interne sont aussi prises en compte.

La méthode des composantes se définit comme suit:

$$P(t + i) = P(t) + [B(t, t + i) - D(t, t + i)] + I(t, t + i) - E(t, t + i) + N(t, t + i). \tag{1}$$

Où, pour une province donnée:

$P(t + i)$ = estimation de la population au temps $t + i$

$P(t)$ = effectifs recensés au temps t

B = nombre de naissances entre t et $t + i$

D = nombre de décès entre t et $t + i$

I = nombre d'immigrants entre t et $t + i$

E = nombre d'émigrants entre t et $t + i$

N = nombre d'immigrants interprovinciaux nets entre t et $t + i$

$(t, t + i)$ = intervalle entre la date du recensement le plus récent et la date de référence de l'estimation.

2.2 Migration interprovinciale

Deux fichiers administratifs sont utilisés pour produire les estimations annuelles et trimestrielles de la migration interprovinciale. Les estimations provisoires sont établies au moyen du fichier des allocations familiales tandis que les estimations définitives le sont à l'aide du fichier d'impôt de Revenu Canada.

2.2.1 Estimations provisoires

L'effectif de migrants adultes est estimé en utilisant les données de migration des enfants dérivées des fichiers des allocations familiales combinées aux rapports des taux d'émigration des adultes à ceux des enfants (f_{jk}) selon le fichier d'impôt de Revenu Canada le plus récent (antérieur d'une année ou deux à la date de référence). Les bénéficiaires d'allocations familiales doivent aviser le ministère de la Santé et du Bien-être social de tout changement d'adresse. Ces changements sont compilés tous les mois, par province d'origine et de destination et selon la taille de la famille (le nombre d'enfants par famille recevant l'allocation). La couverture de la population au moyen des allocations familiales est comparable à celle du recensement (Statistique Canada 1987, p.46). Les estimations du nombre de sortants interprovinciaux pour tous les groupes d'âge sont calculées comme suit:

Utilisation des fichiers administratifs pour estimer la population au Canada¹

RAVI B.P. VERMA et RONALD RABY²

RÉSUMÉ

Ce document étudie l'exactitude des estimations d'émigrants du Canada et de migrants interprovinciaux déduites des fichiers d'allocation familiales et d'impôt de Revenu Canada. L'application de ces fichiers à l'estimation de la population totale du Canada, des provinces et des territoires est évaluée par comparaison aux chiffres du recensement de 1986. On démontre que ces deux fichiers administratifs offrent des séries de données cohérentes et raisonnablement précises sur l'émigration et la migration interprovinciale pour la période de 1981 à 1986. Il en résulte que les estimations de population sont exactes. L'estimation des émigrants produite à partir du fichier des allocations familiales serait plus précise si l'on utilisait le fichier de l'immigration d'Emploi et Immigration Canada dans le calcul du rapport des taux d'émigration des adultes à ceux des enfants.

MOTS CLÉS: Migration interprovinciale; émigration; estimations de population; effectifs recensés; exactitude.

1. INTRODUCTION

Le recensement national, quinquennal depuis 1951, fournit un éventail étendu de données démographiques sur la population canadienne. Cependant, contrairement à quelques autres pays industrialisés, le Canada n'a pas de système d'enregistrement continu de la population, sur lequel baser ses données démographiques et suivre le mouvement des individus entre diverses régions géographiques, les années intercensitaires. Pour combler cette lacune, depuis les années quarante, Statistique Canada a développé un programme d'estimations de la population et des familles. Par exemple, les estimations de la population du Canada, des provinces et territoires, des divisions et régions métropolitaines de recensement sont basées sur les effectifs du dernier recensement ainsi que sur les données provenant de plusieurs fichiers administratifs: les fichiers d'impôt de Revenu Canada et des allocations familiales dans le cas de la migration; les registres de la statistique de l'état civil dans le cas des naissances et des décès et le registre des visas d'immigrant et les fiches relatives au droit d'établissement en ce qui concerne l'immigration.

Les forces et les faiblesses de ces fichiers administratifs pour estimer la population et la migration, déduites de comparaisons au recensement de 1981, ont déjà été décrites antérieurement (Statistique Canada 1987; Verma et Parent 1985; Norris, Britton et Verma 1982). Dans le présent document, l'exactitude des estimations des composantes de l'accroissement démographique des provinces et territoires, basées sur les fichiers des allocations familiales et de Revenu Canada, sera évaluée par comparaison aux données du recensement de 1986. L'évaluation comparera les données de 1971, 1976 et 1981.

Le texte comporte les sections suivantes: l'introduction; les sources de données et méthodes d'estimation; les résultats de l'évaluation; et la conclusion et discussion.

¹ Version révisée d'un document présenté au symposium de Statistique Canada sur les utilisations statistiques des données administratives, novembre 1987.
² Ravi B.P. Verma et Ronald Raby, Division de la démographie, Statistique Canada, 4-A Edifice Jean Talon, Ottawa, Ontario, K1A 0T6.

```
{Initialisation des variables}
i = 1; SamProb = 1; NumRem = SamSize; Gamma = 1/Z[2];
{ Programme pour l'échantillonnage }
while NumRem > 0 do
begin
  if i > 1 and i < PopSize then
    Gamma = Gamma*(1 - z[i]/Z[i] + 1);
    if i = PopSize - NumRem + 1 or Random < Numrem*z[i]/Z[i]
    then
      begin
        if i < > PopSize - NumRem + 1 then
          SamProb = SamProb*NumRem*z[i]/Z[i];
          NumRem = NumRem - 1;
          Sample[SamSize - NumRem, 1] = i;
          Sample[SamSize - NumRem, 2] = SamSize*z[i];
          Sample[SamSize - NumRem, 3] = Gamma;
          end else SamProb = SamProb*(1 - NumRem*z[i]/Z[i]);
          i = i + 1;
        end.
```

BIBLIOGRAPHIE

DREW, J. D., CHOUDHRY, G. H., et GRAY, G. B. (1978). Some methods for updating sample survey frames and their effects on estimation. *Techniques d'enquête*, 4, 225-263.

FELLEGI, I. P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

FELLEGI, I. P. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. *Proceedings of the Section on Social Statistics, American Statistical Association*, 434-442.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size. *Journal of the American Statistical Association*, 58, 183-201.

KISH, L., et SCOTT, A., (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

RAO, J. N. K., HARTLEY, H. O., et COCHRAN, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Série B*, 24, 482-490.

PLATEK, R., et SINGH, M. P. (1978). A strategy for updating continuous surveys. *Metrika*, 25, 1-7.

SUNTER, A. B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.

SUNTER, A. B. (1989). PPS Sampling in multistage designs: does it matter which method? Document soumis à *Journal of Official Statistics*.

Tableau 3
200 Itérations d'une méthode de mise à jour des effectifs $n = 4$, $M = 100$;
effectifs originaux tirés de $R(1,3)$

Cas	Source de π_i	Nombre d'échantillons conservés à l'étape 1	Nombre moyen d'essais à l'étape 2	Valeur P^* estimée
1	$z_{2i} \approx R(1,3)$	134	2.98	0.33
2	$z_{2i} = 2^* z_{1i}$ pour 10% des u.p.é.	153	5.53	0.19
3	$z_{2i} = R(z_{1i}/2, 3z_{1i}/2)$	154	4.17	0.25

Le cas 1, où les nouveaux effectifs sont générés à partir de la même distribution indépendante des valeurs originales, peut être apparentée au scénario de "la pire éventualité". Le cas 2, où 10% des u.p.é. ont un effectif deux fois plus grand qu'avant alors que 90% ont un effectif inchangé, s'approche du scénario du "développement chaotique". Enfin, le cas 3 illustre la perturbation aléatoire des effectifs par une quantité distribuée de façon rectangulaire sur un intervalle équivalent à l'effectif original. Nous pouvons voir d'après le tableau 3 que la probabilité de conserver l'échantillon original varie de 0.67, pour le scénario de "la pire éventualité", à 0.81, pour le scénario du "développement chaotique". Pour ce qui est des cas où l'échantillon original est rejeté, le nombre moyen d'essais requis à l'étape 2 avant d'accepter un nouvel échantillon est très près de la valeur prévue de $1/P^*$.

REMERCIEMENTS

L'auteur tient à remercier le rédacteur en chef ainsi que deux arbitres pour lui avoir communiqué leurs précieux commentaires et pour avoir corrigé les erreurs qui s'étaient glissées dans la version originale.

ANNEXE

Pseudo-code pour la version 1 de la méthode d'échantillonnage progressif avec PPT

On suppose ici que la population des effectifs est déjà classée dans un ordre approprié (ce classement ayant été déterminé, par exemple, au moyen de l'algorithme décrit dans Sunter (1986)) et que l'indice i , en l'occurrence, désigne l'unité. Les effectifs, transformés de manière que leur somme soit égale à 1, sont enregistrés dans un tableau identifié $z[1..PopSize]$ et leurs valeurs cumulatives (depuis $PopSize$ jusqu'à 1), enregistrées dans un tableau identifié $Z[1..PopSize]$. Les noms donnés aux variables illustrent la signification de ces variables. Les résultats sont enregistrés dans un tableau identifié $Sample[1..SampleSize, 1..3]$ et dont les éléments sont, dans l'ordre, l'indice de population i , la probabilité d'échantillonnage des unités π_i et τ_i . La fonction "Random" produit un nombre aléatoire distribué selon une loi uniforme sur l'intervalle (0,1). Les décalages qui paraissent dans le code ci-dessous ont pour but de mettre en relief les instructions composées.

Tableau 1
Probabilités d'échantillonnage

UPÉ	z_{1i}	z_{2i}
1	0.15	0.35
2	0.20	0.30
3	0.30	0.20
4	0.35	0.15

Tableau 2

(1) Echantillon	(2) $P_1(S)$	(3) $P_2(S)$	(4) $P_{2 1}(S)$	(5) $P_{2 2}(S)$
1,2	0.0231	0.3231	1.0	0.9286
1,3	0.1154	0.2154	1.0	0.4643
1,4	0.1615	0.1615	1.0	0
2,3	0.1615	0.1615	1.0	0
2,4	0.2154	0.1154	0.5357	0
3,4	0.3231	0.0231	0.0715	0

On peut vérifier que la probabilité totale de conserver le même échantillon, laquelle équivaut à la somme des produits des valeurs indiquées dans les colonnes (2) et (4), est 0.5465. En comparant cette valeur à la probabilité totale de conserver l'échantillon original si le nouvel échantillon est prélevé de façon indépendante, c'est-à-dire $\sum_i P_1(S_i)P_2(S_i) = 0.1168$, nous constatons que nous avons pu accroître considérablement la probabilité de conserver le même échantillon.

3.2 Exemple 2

Nous allons maintenant prendre un exemple plus réaliste. Prélevons $n = 4$ u.p.é. parmi 100; les effectifs "originaux" sont tirés (de façon indépendante) d'une distribution uniforme ou rectangulaire $R(1,3)$. Les "nouveaux" effectifs sont déterminés par diverses méthodes que nous décrivons plus bas. Avec cet exemple, nous ne sommes plus en mesure d'énumérer tous les échantillons possibles ni d'exécuter manuellement l'échantillonnage et le calcul des probabilités de sélection. En revanche, nous n'avons eu aucune difficulté à écrire un programme d'ordinateur qui exécuterait ces tâches et appliquerait la méthode de ré-échantillonnage. Grâce à ce programme, nous avons pu exécuter, pour chaque exemple, 200 itérations d'un échantillonnage selon la version 2 de Sunter avec probabilités proportionnelles à la première série d'effectifs, puis 200 itérations d'un ré-échantillonnage avec probabilités proportionnelles à la deuxième série d'effectifs. Le programme, exécuté sur un micro-ordinateur compatible avec un modèle XT fonctionnant à 7.16 MHz, a produit et trié les populations d'effectifs et exécuté les 200 échantillonnages et les 200 ré-échantillonnages en trois minutes environ.

Version 2: Classer les éléments de la population de telle manière que

- (a) $n z_i \leq Z_i; i = 1, 2, \dots, N - n - 1$
- (b) $(n - i) z_i < Z_i; i \geq i \geq N - n.$

Ensuite

- (i) prélever des unités selon la formule $P(U_i | n_i) = n z_i / Z_i$ jusqu'à ce que $n_i = 0$ ou $n_i = N - i,$

- (ii) si $n_i > 0$, éliminer une des unités qui restent, par exemple celle qui a pour indice j , avec une probabilité $1 - n_i z_j / Z_i$ et prélever les autres.

Sunter (1986) décrit un algorithme permettant de déterminer un mode de classement qui satisfasse aux conditions de la seconde version; cet algorithme a été incorporé au programme qui a servi à la simulation dans les exemples ci-dessous. Selon les deux versions de la méthode, on peut calculer π_{ij} à l'aide de la formule

$$\pi_{ij} = n(n - 1) z_i z_j T_{ij},$$

où $i < j$ (en ce qui concerne le classement utilisé) et

$$T_1 = 1/Z_2$$

$$T_i = (1/Z_i + 1)(1 - z_1/Z_2) \dots (1 - z_{i-1}/Z_i).$$

Ces expressions sont exactes pour $i < j \leq N - n + 1$ et constituent, dans d'autres cas, une très bonne approximation. Elles sont faciles à résoudre et permettent de calculer une estimation de la variance avec un biais négligeable, ce qu'aucune autre méthode PPTSR avec $n > 2$ ne peut offrir.

Nous avons reproduit en annexe un pseudo-code de type Pascal pour un programme qui préleve un échantillon selon la version 1 et qui calcule du même coup la probabilité de sélection de cet échantillon et la valeur de T_i pour chaque unité prélevée. Le programme peut facilement être appliqué à la version 2 ou modifié de manière à permettre le calcul de $P(S)$ pour un échantillon déjà prélevé.

3.1 Exemple 1

Pour illustrer les méthodes exposées ci-dessus, nous allons tout d'abord utiliser un exemple où $n = 2$ et $N = 4$ ce qui permettra un calcul manuel et une énumération de tous les échantillons possibles. On remarquera que pour obtenir les "nouvelles" tailles de population, nous avons simplement inversé l'ordre des probabilités d'échantillonnage initiales. L'algorithme de classement de la version 2 donne (4, 1, 2, 3) pour la première série d'effectifs et (1, 4, 3, 2) pour la seconde série. Il y a six échantillons possibles, lesquels figurent dans la colonne (1) du tableau 2, et leurs probabilités de sélection respectives selon l'algorithme de la version 2 sont facilement calculables. Ces probabilités figurent dans les colonnes (2) et (3) du tableau 2. La colonne (4) indique la probabilité de conserver un échantillon à l'étape 1, étant donné qu'il s'agit de l'échantillon original tandis que la colonne (5) indique la probabilité conditionnelle qu'un échantillon soit accepté à n'importe quel stade de l'étape 2, étant donné qu'il n'a été accepté à aucun stade antérieur.

3. APPLICATION ET EXEMPLES

Le nouveau plan d'échantillonnage peut différer de l'ancien à d'autres points de vue que celui de la probabilité d'échantillonnage des unités. Nous pourrions passer, par exemple, d'un échantillonnage systématique avec PPT à un échantillonnage progressif avec PPT (Sunter 1986, 1989) ou même d'un échantillonnage avec PPTAR (probabilité proportionnelle à la taille avec remise) à un échantillonnage avec PPTSR. Dans ce dernier cas toutefois, un échantillon initial qui contiendrait plus d'une fois la même u.p.é. aurait une probabilité de sélection nulle selon le plan PPTSR. Notons que la méthode peut encore être appliquée lorsque de nouvelles u.p.é. sont incluses dans la strate mais que la taille de l'échantillon demeure la même. La méthode a probablement sa plus grande utilité (du point de vue de la probabilité de conserver le même échantillon) lorsque l'ancien et le nouveau plan d'échantillonnage sont tels que la totalité ou la presque totalité des échantillons sont réalisables et que leurs probabilités de sélection sont approximativement proportionnelles au produit des probabilités d'échantillon-nage de leurs unités respectives. Dans ces conditions, et dans la mesure où la variation des effets n'est pas trop grande, $P_1(S_i)$ et $P_2(S_i)$ auront tendance à être comparables de sorte que la probabilité de conserver le même échantillon sera relativement levée. La méthode mention-née plus haut a toutes les propriétés voulues pour ce genre d'application. Comme c'est cette méthode que nous utiliserons dans les exemples présentés plus loin, nous allons maintenant nous attacher à la décrire. Cette méthode existe en deux versions; dans les deux cas, il s'agit de déterminer un classement approprié pour les éléments de la population et de disposer les tailles de population (dont la somme est supposée égale à 1) dans l'ordre inverse (pour ainsi dire) de manière à obtenir:

$$Z_i = \sum_N^i z_j, i = 1, 2, \dots, N.$$

Version 1: Classer les éléments de la population de telle manière que

- (a) $nz_i \leq Z_i; i = 1, 2, \dots, N - n$
- (b) $(n - i)z_i > Z_i; i = n, n + 1, \dots, N - 1.$

Ensuite, prélever exactement n unités suivant la formule:

$$P(U_i | n_i) = \begin{cases} 1 & \text{si } n_i = N - i + 1 \\ n_i z_i / Z_i & \text{dans le cas contraire} \end{cases}$$

où n_i est le nombre d'unités qu'il faut encore prélever lorsque nous arrivons à la i -ième unité de population.

Il est toujours possible de satisfaire aux conditions (a) et (b). Par exemple, les deux condi-tions sont respectées lorsque, évidemment, on classe les éléments par ordre croissant de taille ou lorsqu'on les classe par ordre décroissant jusqu'à ce que (b) ne soit plus respectée (le cas échéant), puis par ordre croissant. Un avantage de ce mode de classement est qu'il tend à com-penser les effets mineurs (sinon négligeables) d'une dérogation à la règle de la PPT pour les n dernières unités (voir Sunter 1986). La seconde version de la méthode de mise à jour élimine toute possibilité de dérogation à la règle de la PPT en mettant en application le principe sui-vant: s'il reste $n_i + 1$ unités dans la population (pour tout i), il est normalement possible d'exclure l'une de ces unités avec une probabilité appropriée et de conserver les autres.

$$P^* = \sum_{i:P_2(S_i) > P_1(S_i)} (1 - P_2(S_i)/P_1(S_i))P_1(S_i) \\ = \sum_{i:P_2(S_i) > P_1(S_i)} (P_1(S_i) - P_2(S_i)) \tag{1}$$

où i sert désormais d'indice pour les sous-ensembles n -tuple des N unités de la population, et

$$P^{**} = 1 - \sum_{i:P_1(S_i) > P_2(S_i)} (1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$= 1 - \sum_{i:P_1(S_i) > P_2(S_i)} (P_2(S_i) - P_1(S_i)) \tag{2}$$

Comme $\sum_i P_1(S_i) = \sum_i P_2(S_i) = 1$, il est facile de constater que les termes de sommation des équations (1) et (2) doivent être égaux entre eux et que, par conséquent, $P^* = 1 - P^{**}$. Si nous désignons la probabilité d'échantillonnage ultime par P' , nous avons (selon un plan):

$$\text{Pour } i: P_2(S_i) < P_1(S_i)$$

$$P'(S_i) = P_1(S_i) (P_2(S_i)/P_1(S_i))$$

$$= P_2(S_i), \text{ ce qui est le résultat recherché.}$$

$$\text{Pour } i: P_2(S_i) \geq P_1(S_i)$$

$$P'(S_i) = P_1(S_i) + P^*(1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$+ P^*P^{**}(1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$+ P^*(P^{**})^2(1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$+ P^*(P^{**})^3(1 - P_1(S_i)/P_2(S_i))P_2(S_i)$$

$$+ \dots$$

$$= P_1(S_i) + P^*(P_2(S_i) - P_1(S_i))/(1 - P^{**})$$

$$= P_2(S_i)$$

ce qui est le résultat recherché.

Enfin, nous remarquons que le nombre espéré d'essais à l'étape 2, étant donné que l'échantillon original n'a pas été conservé à l'étape 1, est défini par la formule de la distribution binomiale pour le temps d'attente $1/(1 - P^{**}) = 1/P^*$.

formellement à n'importe quelle méthode de type PPTSR pour laquelle il est possible d'établir la probabilité de sélection de n'importe quel échantillon, cette méthode de mise à jour est surtout utile pour les méthodes PPTSR où la totalité, ou presque, des sous-ensembles n -uple sont des échantillons possibles avec probabilité approximativement proportionnelle au produit des probabilités d'échantillonnage de leurs unités respectives. À des fins d'illustration, nous allons utiliser une méthode de ce genre, à savoir l'échantillonnage progressif avec ppt (Sunter 1986, 1989).

2. FONDEMENTS THÉORIQUES DE LA MÉTHODE DE MISE À JOUR

Soit un échantillon PPTSR tiré initialement avec probabilités $\{\pi_{11}, \pi_{12}, \dots, \pi_{1n}\}$ proportionnelles aux effectifs originaux $\{z_{11}, z_{12}, \dots, z_{1n}\}$. Nous voulons tirer un nouvel échantillon PPTSR suivant une nouvelle série de probabilités $\{\pi_{21}, \pi_{22}, \dots, \pi_{2n}\}$ proportionnelles aux effectifs révisés $\{z_{21}, z_{22}, \dots, z_{2n}\}$. Cependant, nous voulons procéder de telle manière qu'il y ait de fortes chances de conserver l'échantillon original.

Nous supposons que, pour n'importe quel n -tuple S , y compris bien sûr S' , l'échantillon initial, il est possible de calculer $P_1(S)$, la probabilité d'échantillonnage selon le plan original, et $P_2(S)$, la probabilité d'échantillonnage selon un nouveau plan. Pour de nombreux échantillons, l'une ou l'autre de ces probabilités ou les deux pourront être nulles selon de nombreux plans (p. ex.: échantillonnage systématique avec ppt); il est clair, toutefois, que $P_1(S')$ ne peut être nul en l'occurrence.

Le processus est le suivant:

- Etape 1: a) Calculer $P_1(S'), P_2(S')$.
- b) Si $P_2(S') \geq P_1(S')$, conserver l'échantillon.
- c) Si $P_2(S') < P_1(S')$, conserver l'échantillon avec une probabilité $P_2(S')/P_1(S')$. Si rejeté, passer à l'étape 2.
- Etape 2: a) Si l'échantillon original n'a pas été conservé, tirer un nouvel échantillon, appelé S_1 avec un probabilité $P_2(S_1)$. Si $P_2(S_1) < P_1(S_1)$ rejeter l'échantillon; autrement, conserver l'échantillon avec une probabilité $1 - P_1(S_1)/P_2(S_1)$. Si rejeté, passer à l'étape 2(b).
- b) Si l'échantillon de l'étape 2(a) a été rejeté, tirer un nouvel échantillon et procéder comme à l'étape, 2(a).
- c), d), ... Reprendre les étapes 2(a), 2(b), ... jusqu'à ce qu'un échantillon soit accepté.

L'échantillon que l'on obtient éventuellement par ce processus a la structure de probabilité voulue tant pour les probabilités d'échantillonnage des unités que pour les probabilités composées d'échantillonnage des paires d'unités. Autrement dit, on peut considérer cet échantillon comme prélevé suivant le nouveau plan. En particulier, puisqu'il a la même structure de probabilité composée, il a la même variance d'échantillonnage.

Définissons P^* comme la probabilité que le processus ne se termine pas à l'étape 1 et P^{**} comme la probabilité conditionnelle qu'il ne se termine pas à l'étape 2(a), étant donné qu'il n'a pas pris fin à l'étape 1. Evidemment, P^{**} représente aussi la probabilité conditionnelle que le processus ne se termine pas à une étape subséquente quelconque, étant donné qu'il n'a pas pris fin à n'importe quelle autre étape qui la précède. Nous avons alors

Mise à jour de la taille de population dans un plan PPTSR

ALAN SUNTER¹

RÉSUMÉ

On doit parfois mettre à jour l'échantillon PPTSR des unités de premier degré (u.p.é.) d'un plan de sondage à plusieurs degrés afin de tenir compte des nouveaux effets observés dans ces unités. Toutefois, compte tenu des ressources considérables que nécessitent l'établissement des cartes dans les u.p.é., la segmentation, le listage, le recrutement des recenseurs, etc., on voudrait conserver dans la mesure du possible les mêmes u.p.é., pourvu que les probabilités d'échantillonnage puissent être considérées comme proportionnelles aux nouveaux effets. La méthode exposée dans cet article diffère de toutes celles qui ont été décrites antérieurement en ce qu'elle vaut pour n'importe quelle taille d'échantillon et qu'elle ne nécessite pas une énumération de tous les échantillons possibles. De plus, il n'est pas nécessaire de conserver la même méthode d'échantillonnage. Par conséquent, la méthode proposée ici rend possible non seulement la mise à jour de la taille de population mais aussi l'utilisation d'une nouvelle méthode d'échantillonnage.

MOTS CLÉS: PPTSR; mise à jour de l'échantillon; échantillonnage progressif avec PPT.

1. INTRODUCTION

On doit parfois mettre à jour l'échantillon PPTSR des unités de premier degré (u.p.é.) d'un plan de sondage à plusieurs degrés afin de tenir compte des nouveaux effets observés dans ces unités. Par exemple, ce genre d'opération devient nécessaire lorsque les u.p.é. sont des secteurs de population et du logement pour ces secteurs sont rendus publics par suite d'un recensement ou lorsqu'on décide de faire une mise à jour provisoire de l'effectif d'une strate d'échantillonnage après avoir observé une croissance inégale des effets des SD durant une période intercensitaire. Toutefois, compte tenu des ressources considérables que nécessitent l'établissement des cartes dans les u.p.é., la segmentation, le listage, le recrutement des recenseurs, etc., nous aimerions conserver dans la mesure du possible les mêmes u.p.é., pourvu que les probabilités d'échantillonnage, qui étaient initialement proportionnelles à l'ancien effectif, soient désormais proportionnelles au nouvel effectif. Kish et Scott (1971) font une analyse détaillée de la question pour $n = 1$, cette analyse étant la généralisation d'une méthode qui avait été élaborée antérieurement par Keyfitz (1951). Ils précisent que leur méthode peut s'appliquer sans problème à l'échantillonnage avec remise (PPTAR) pour $n > 1$. Elle peut aussi être utilisée (Drew, Choudhry et Gray 1978; Platek et Singh 1978) pour les plans PPTSR où $n > 1$ lorsque la méthode utilisée est celle de Rao, Hartley et Cochran (1962), laquelle prévoit la formation de n groupes aléatoires et le prélèvement d'une u.p.é. dans chaque groupe. Par contre, elle devient inapplicable si, comme il faut s'y attendre, nous désirons former de nouveaux groupes aléatoires en fonction des effets révisés. Fellegi (1966) décrit deux méthodes pouvant s'appliquer à un échantillon PPTSR de taille $n = 2$ prélevé au moyen de la méthode élaborée quelques années plus tôt (Fellegi 1963).

La méthode décrite ci-dessous s'apparente à la seconde méthode de Fellegi et donne des résultats qui se rapprochent sensiblement de ceux de Fellegi lorsqu'elle est appliquée aux mêmes exemples. En revanche, elle ne nécessite pas une énumération de tous les échantillons possibles et peut donc s'appliquer à n'importe quelle valeur de n et N . Bien qu'elle puisse s'appliquer

¹ Alan Sunter, Président, A.B. Sunter Research Design & Analysis Inc., 63, 5e Av., Ottawa, Canada, K1S 2M3.

- LEONARD, K. J. (1988). Credit scoring via linear logistic models with random parameters. Thèse de doctorat, Département of Decision Sciences and Management Information Systems, Concordia University, Montréal.
- LEVY, P. S., (1971). The use of mortality data in evaluating synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 323-331.
- LEVY, P. S., et FRENCH, D. K. (1978). Estimation of health characteristics. *Vital and Health Statistics*, Ser. 2, No. 75, NCHS, Washington, DC.
- MADOW, W. G., et HANSEN, M. H. (1975) On statistical models and estimation in sample surveys. Dans *Contributed Papers*, 40th Session of the International Statistical Institute, Warsaw, Poland, 554-557.
- MIAO, L. L. (1977). An empirical Bayes approach to analysis of inter-area variation. Thèse de doctorat, Department of Statistics, Harvard University.
- MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-54.
- O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.
- PLATEK, R., et SINGH, M. P. (1986). *Small Area Statistics—An International Symposium '85* (Contributed Papers). Dans Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University—Université d'Ottawa, Canada.
- PURCELL, N. J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- PURCELL, N. J., et KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.
- ROBERTS, G., RAO, J. N. K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROBBINS, H. I. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium*. Berkeley: University of California Press, 15.17-164.
- ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALL, R. M. (1973). Discussion of papers by Gonzalez and Eriksen. *Proceedings of the Section on Social Statistics, American Statistical Association*, 42-43.
- SÄRNDA, C. E. (1984). Design consistent versus model dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCHABLE, W. L. (1979). A composite estimator for small area statistics. Dans *Synthetic Estimates for Small Areas* (NIDA Research Monograph 24), édité par J. Steinberg. Rockville, MD: National Institute on Drug Abuse, 36-53.
- STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. Dans *Small Area Statistics*, (Eds. R. Platek, J. N. K. Rao, C. E. Särndal et M. P. Singh.). New York: Wiley, 124-137.
- TOMBERLIN, T. J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.
- U.S. National Center for Public Health Statistics (1968). *Synthetic State Estimates of Disability*, PHS Publication No. 1759.
- WEISBERG, H. I., TOMBERLIN, T. J., et CHATTERJEE, S. (1984). Predicting insurance losses under a cross-classification: a comparison of alternative approaches. *Journal of Business and Economic Statistics*, 2, 170-178.
- WONG, G. Y., et MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

BIBLIOGRAPHIE

BATTESE, G. E., HARTER, R. M., et FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

CRESSIE, N. (1988). Dans quelles circonstances les opérations de redressement améliorent-elles les chiffres du recensement? *Techniques d'enquête*, 14, 191-208.

DEMING, W. E., et STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMPSTER, A. P., LAIRD, N. M., et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (avec discussion). *Journal of The Royal Statistical Society, Sér. B*, 39, 1-38.

DEMPSTER, A. P., RUBIN, D. B., et TSUTAKAWA, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.

DEMPSTER, A. P., et TOMBERLIN, T. J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, 88-94.

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography* 10, 137-159.

ERICKSEN, E. P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

ERICKSEN, E. P. (1980). Can regression be used to estimate local undercount adjustments? *Proceedings of the Conference on Census Undercount*, 55-61.

EFRON, B., et MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.

FAY, R. E., et HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 33-36.

GONZALEZ, M. E., et HOZA, C. (1976). Small area estimation of unemployment. *Proceedings of the Section on Social Statistics, American Statistical Association*, 437-443.

GONZALEZ, M. E., et HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

GONZALEZ, M. E., et WAKSBERG, J. L. (1975). Estimation of the error of synthetic estimates. Articles non publiés présentés à la première conférence de l'Association internationale des statisticiens d'enquête à Vienne.

HABERMAN, S. J. (1978). *Analysis of Qualitative Data Volume I: Introductory Topics*. New York: Academic Press.

HOLT, D., SMITH, T. M. F., et TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

JAMES, W., et STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 361-379.

LAKE, P. (1979). A predictive approach to subdomain estimation in finite populations. *Journal of the American Statistical Association*, 74, 355-358.

LAIRD, N. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.

LAIRD, N., et LOUIS, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (avec discussion). *Journal of the American Statistical Association*, 82, 739-757.

5. CONCLUSIONS

Par la simulation d'un échantillon à deux degrés, où les UPE sont des sous-régions, nous avons pu constater que l'estimateur empirique de Bayes modifié était généralement supérieur à deux autres estimateurs courants. L'évaluation des estimateurs a porté sur trois aspects: biais, erreur type et utilité de mesures de variation estimables. La simulation montre que l'estimateur classique est supérieur aux deux autres au point de vue du biais, ce qui est tout à fait normal étant donné qu'il est non biaisé selon le plan. En outre, on peut calculer des estimations valables de la précision P pour cet estimateur en se servant de méthodes standard. Cependant, ces estimateurs sont caractérisés par un P élevé du fait qu'ils sont construits à partir d'une quantité limitée de données. En effet, contrairement aux deux autres estimateurs, l'estimateur classique ne permet pas d'établir des estimations pour les sous-régions qui ne sont pas incluses dans l'échantillon.

À l'autre bout du spectre, l'estimateur synthétique est beaucoup plus stable que les deux autres. Comme toutes les estimations reposent sur des données provenant de tout l'échantillon, la variance d'échantillonnage est beaucoup plus faible que celle des deux autres estimateurs. Par contre, l'estimateur synthétique n'est pas conçu pour des sous-régions qui sont très différentes des autres même lorsqu'il existe des données d'échantillon qui tendent à confirmer cet état de fait. En outre, les estimations de variabilité pour cet estimateur, sous forme d'écart type d'échantillonnage, sont particulièrement trompeuses puisqu'elles ne tiennent compte d'aucune entorse au modèle à effets fixes.

Constituant une solution intermédiaire, l'estimateur empirique de Bayes modifié s'avère acceptable sur les trois plans étudiés. Du fait qu'il utilise les données de sous-régions (dans la mesure où il s'agit de données fiables), cet estimateur est à l'abri des biais élevés qui caractérisent l'estimateur synthétique. Par ailleurs, en intégrant des données recueillies dans tout l'échantillon, cet estimateur a une variance d'échantillonnage et, règle générale, un P moins élevés que ceux de l'estimateur sans biais. Enfin, les variances a posteriori sont des mesures de variabilité utiles dans ce cas.

Il y a encore beaucoup à faire dans l'analyse des estimateurs proposés. Premièrement, il faudrait voir quelles seraient les conséquences de l'utilisation de véritables estimateurs empiriques de Bayes au lieu d'estimateurs modifiés. Il faudrait aussi définir des lignes directrices à propos du nombre minimum d'unités d'échantillonnage requises pour garantir la validité de l'inférence bayésienne. Comme les vrais estimateurs empiriques de Bayes utilisent des variances a priori estimées, il faut des méthodes qui puissent tenir compte de cette source de variation additionnelle. On pourrait utiliser, par exemple, les techniques "bootstrap" analysées par Laird et Louis (1987). En deuxième lieu, il faudrait généraliser les méthodes d'estimation de manière qu'elles puissent s'appliquer à des plans d'échantillonnage à trois degrés ou plus. Même si cela est facilement concevable au point de vue théorique, il en est tout autrement au point de vue statistique. Enfin, il faudrait appliquer ces méthodes à des données réelles avant d'en recommander officiellement l'utilisation pour l'établissement d'estimations régionales.

REMERCIEMENTS

Les auteurs tiennent à manifester leur gratitude pour les précieux commentaires que leur ont fournis un rédacteur associé et un arbitre, et le soutien financier que leur ont apporté le Conseil de recherches en sciences naturelles et en génie du Canada et l'Université Concordia.

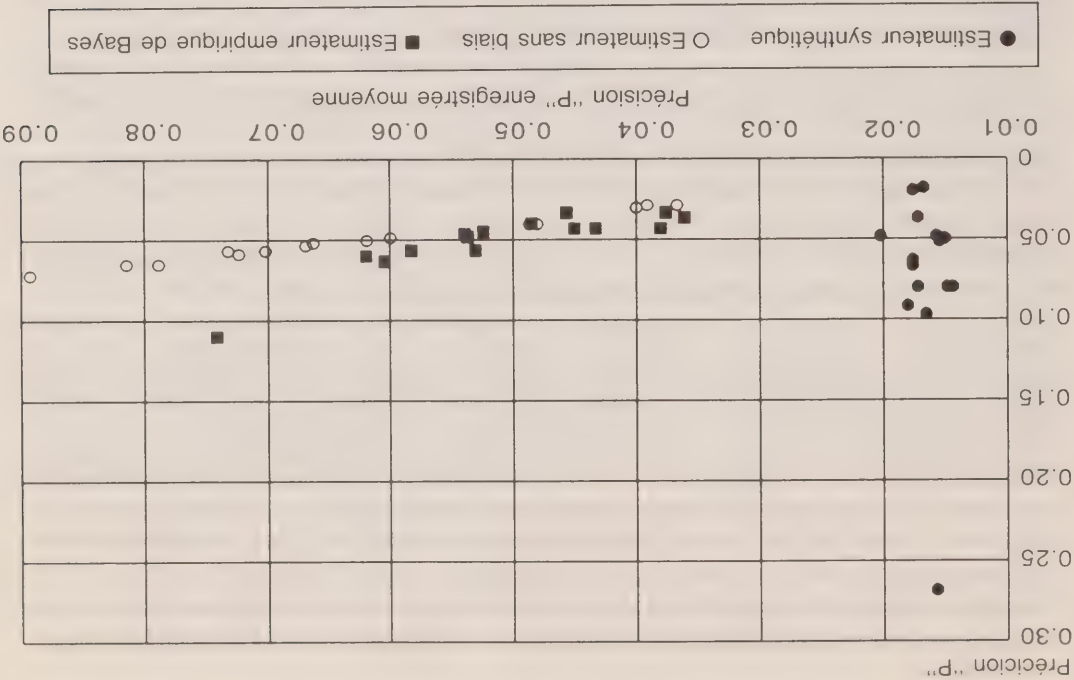


Figure 3. Précision "P" empirique en fonction de la précision P enregistrée, pour chacune des trois méthodes d'estimation

horizontal correspond au "P" enregistré. Pour ce qui est de l'estimateur non biaisé classique, le "P" enregistré correspond simplement à l'écart type d'échantillonnage pour l'échantillonnage aléatoire simple. Pour l'estimateur synthétique, il s'agit aussi de l'écart type d'échantillonnage, corrigé en fonction de l'échantillonnage en grappes. En ce qui a trait à l'estimateur empirique de Bayes, le "P" enregistré est la racine carrée de la variance a posteriori de la proportion estimée, établie à l'aide des méthodes décrites dans la section 3.2.

On remarquera tout d'abord que les points que se rapportent à l'estimateur sans biais sont disposés comme s'ils se trouvaient sur une droite, indiquant pas le fait même une très grande similitude entre les P enregistrés et les P empiriques. Ce résultat est peu surprenant étant donné l'absence de biais dans ce cas; les P enregistrés et les P empiriques sont donc simplement des écarts types d'échantillonnage. Par contre, les points qui se rapportent à l'estimateur synthétique sont concentrés dans l'intervalle [0.015, 0.020] sur l'axe horizontal. Dans ce cas, les P enregistrés sont des estimations d'écart type d'échantillonnage, qui sont toutes très faibles étant donné qu'il s'agit d'estimateur synthétique. Toutefois, si l'on examine les P empiriques, on constate une toute autre répartition. Les P empiriques vont de 0.015 à 0.100 et il y a une valeur aberrante au-delà de 0.250 (sous-région 9). Dans ce cas, les variances d'échantillonnage ne suffisent pas pour mesurer la variabilité des estimations.

L'estimateur empirique de Bayes modifié constitue une fois de plus une solution intermédiaire. Toutefois, pour ce qui a trait au rapport entre le P enregistré et le P empirique, l'estimateur sans biais. À l'exception du point qui a trait à la sous-région 9, les P enregistrés moyen sont voisins des P empiriques correspondantes.

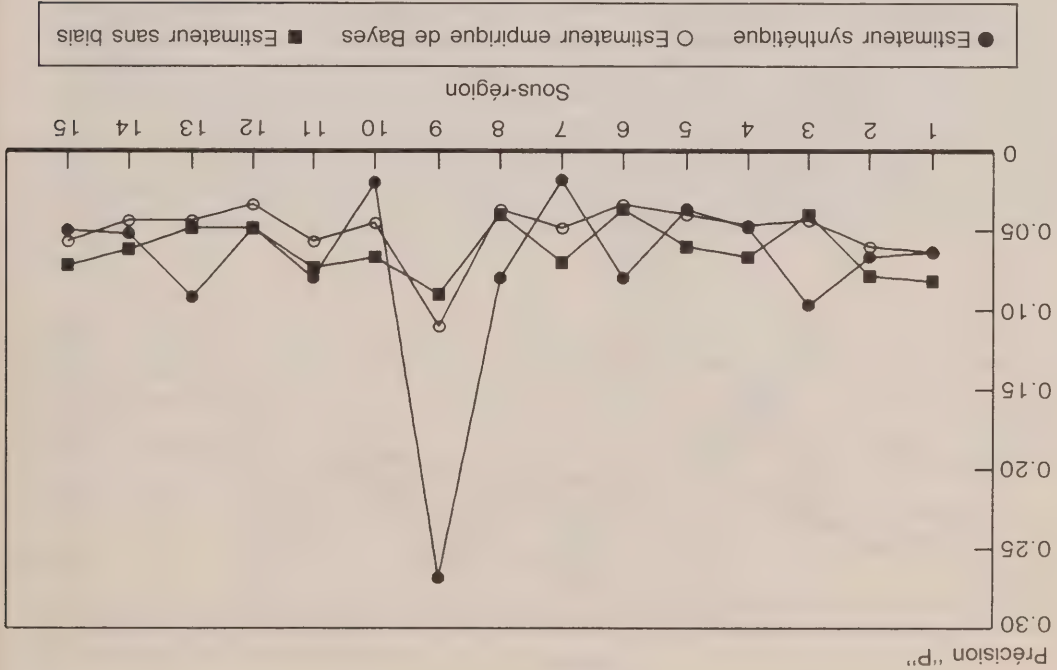


Figure 2. Précision "P" empirique des trois estimateurs par sous-région

synthétique est aussi très élevé à cause du biais élevé de l'estimateur. Quant à l'estimateur empirique de Bayes modifié, le fait qu'il regroupe de toutes les sous-régions lui permet d'afficher un P relativement moins biaisé sans être entaché d'un biais comme celui qui caractérise l'estimateur synthétique dans les sous-régions où l'effet de sous-région est prononcé. Dans tous les cas sauf deux, l'estimateur empirique de Bayes modifié a une valeur de P moindre que l'estimateur sans biais. Pour la sous-région 3, les deux estimateurs ont à peu près le même P tandis que pour la sous-région 9, où l'effet de sous-région est prononcé, l'estimateur empirique de Bayes modifié a une valeur de P légèrement plus élevée que celle de l'estimateur sans biais. En résumé, l'estimateur empirique de Bayes modifié est parfois le plus efficace des trois mais jamais le moins efficace.

Une des principales carences des estimateurs synthétiques du modèle à effets fixes est la difficulté qu'on a d'obtenir des mesures de précision valables. Les seules mesures qu'il est possible d'obtenir sont les variances d'échantillonnage. Il n'existe pas de mesure, comme le biais, qui mette en évidence les faiblesses du modèle. En revanche, pour ce qui a trait aux estimateurs sans biais, les estimations usuelles de la variabilité d'échantillonnage sont aussi les estimations de l'erreur quadratique moyenne (EQM) puisque il n'y a pas de biais. Enfin pour ce qui concerne les estimateurs empiriques de Bayes, on se réfère à la matrice des covariances a posteriori des paramètres pour obtenir des mesures de variabilité. Ces variances a posteriori indiquent la variabilité d'échantillonnage ainsi que le "biais" qui découle des faiblesses du modèle à effets fixes simple. Cette source de variation est mesurée par la variabilité des paramètres d'effets de sous-région.

Dans la figure 3, nous comparons l'utilité de ces mesures de variation. Sur l'axe vertical, nous trouvons le P empirique, que l'on obtient en comparant, pour chaque sous-région, les estimations des échantillons répétées avec les proportions de population connues. L'axe

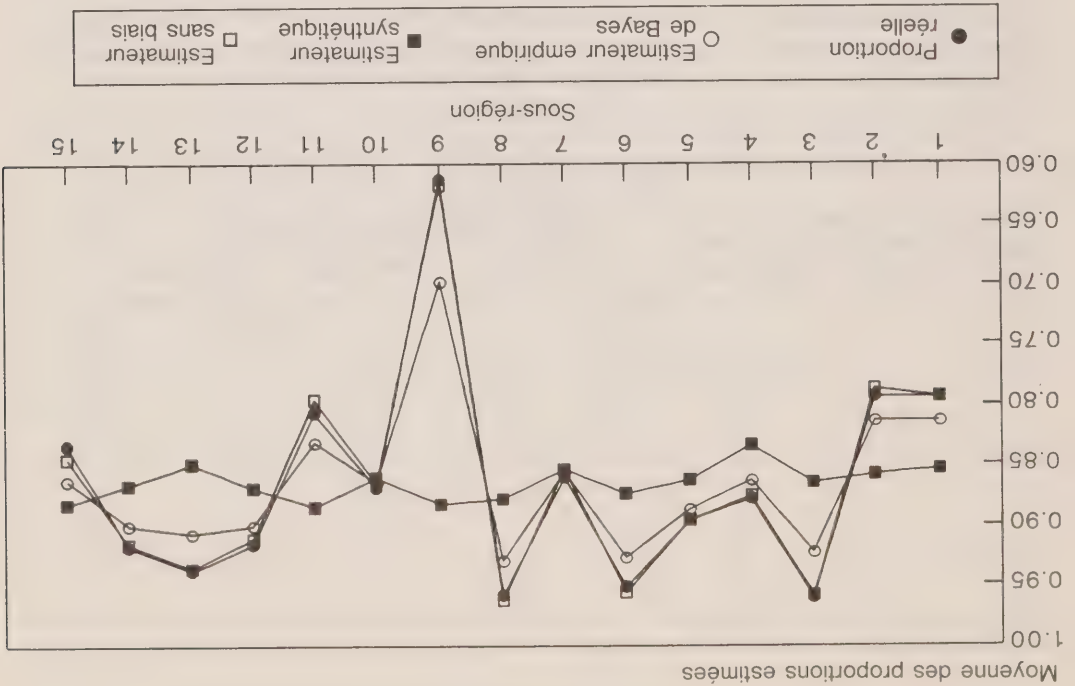


Figure 1. Moyennes des taux d'activité estimés établies selon les trois méthodes d'estimation pour chaque sous-région

de la population sont représentés dans le graphique comme des "proportions réelles". Ces taux coïncident presque parfaitement avec les estimations non biaisées moyennes, ce qui explique qu'ils ne paraissent pas sur le graphique pour la plupart des sous-régions. Cela confirme la propriété d'être sans biais des estimateurs classiques.

Les estimations synthétiques varient peu d'une sous-région à l'autre. Comme les taux d'activité pour les 15 sous-régions reposent sur les mêmes paramètres estimés fixes, la seule source de variation entre les sous-régions est la faible variabilité des distributions réalisées des variables explicatives. Le biais peut être élevé, comme c'est le cas pour la sous-région 9, où la méthode synthétique engendre un biais positif élevé. En revanche, précisons qu'il ne fallait pas s'attendre à ce que cette méthode soit très efficace à cause de la faible variabilité des distributions des variables explicatives d'une sous-région à l'autre.

Les estimations moyennes calculées à l'aide de l'estimateur empirique de Bayes se situent entre les estimations non biaisées et les estimations synthétiques. Elles sont biaisées (au sens classique du terme) mais leur biais est moindre que celui de l'estimateur synthétique du modèle à effets fixes.

Nous avons aussi calculé la précision P (racine carrée de l'EQM) (ET) empirique des trois estimateurs; celles-ci sont représentées dans la figure 2. Par ce graphique, nous pouvons voir ce qui est des sous-régions 7 et 10, où l'effet de sous-région est presque nul, l'espérance mathématique de l'estimateur synthétique est voisine de la proportion de la population. Dans ces sous-régions, l'estimateur synthétique est celui des trois qui a, de loin, la valeur de P l'ET la moins élevée. Etant construit à l'aide de données prises dans tout l'échantillon, l'estimateur synthétique est caractérisé par une faible variance d'échantillonnage. En revanche, pour ce qui a trait à la sous-région 9, où l'effet de sous-région est très prononcé, le P de l'estimateur

Tableau 1

Taux d'activité de la population par sous-région

Sous-région	1	2	3	4	5	6	7	8
Taux d'activité	0.79	0.79	0.96	0.88	0.90	0.95	0.86	0.96
Sous-région	9	10	11	12	13	14	15	
Taux d'activité	0.61	0.87	0.81	0.91	0.94	0.92	0.83	

Dans l'expression (4.1), θ_1 et θ_2 représentent les effets fixes relatifs aux hommes et aux femmes respectivement. En d'autres termes, le risque relatif pour l'activité chez les hommes est $\exp [0.5] = 1.65$. Le paramètre β désigne la pente et est lié à l'âge tandis que ϕ_i représente les effets aléatoires logistiques qui se rattachent aux 15 UPE ou sous-régions.

Les variables explicatives ont été générées avec des distributions identiques dans les 15 sous-régions. L'âge était distribué uniformément sur l'intervalle [20, 40], le sexe de chaque personne était déterminé à l'aide d'une distribution de Bernoulli avec une proportion de 0.5, et les deux variables explicatives étaient supposées être distribuées de façon indépendante. Le tableau 1 donne le taux d'activité de la population pour les 15 sous-régions. Comme nous avons supposé que la distribution des variables explicatives était la même dans les 15 sous-régions, les effets aléatoires ϕ_i constituaient donc la seule source de variation entre les sous-régions. Le caractère aléatoire de ces effets peut entraîner une variation notable du taux d'activité entre les sous-régions, comme l'illustre particulièrement bien la sous-région 9.

Les proportions observées pour les échantillons de sous-régions ont servi d'estimations non biaisées. L'estimateur synthétique était fondé sur le modèle logit à effets fixes ci-dessous,

$$\text{logit} (\pi_{\mu\nu}) = \theta_{\mu} \tag{4.2}$$

où $\pi_{\mu\nu}$ et θ_{μ} sont définis comme dans le modèle à effets aléatoires (2.3). Notons que l'estimateur sans biais est formé à l'aide des données d'une seule sous-région tandis que l'estimateur synthétique est formé à l'aide de données provenant de toutes les sous-régions. En revanche, le biais des estimateurs synthétiques variera en fonction de l'incapacité du modèle (4.2) de tenir compte des différences entre les sous-régions.

Le troisième estimateur analysé ici est une version modifiée de l'estimateur empirique de Bayes proposé dans la section 3. Compte tenu du temps que prend l'ordinateur pour estimer la composante de la variance liée aux effets de sous-région, on s'est servi en fait de l'estimateur de Bayes décrit à la section 3.1. La variance a priori utilisée dans les circonstances était la valeur connue de la variance définie en (4.1), qui a servi à la simulation des données. À cause de ce compromis, les résultats indiqués ci-dessous pour l'estimateur "empirique de Bayes" devraient être légèrement meilleurs que ceux que l'on obtiendrait au moyen d'un véritable estimateur empirique de Bayes. Toutefois, d'après des analyses de sensibilité visant à déterminer l'effet de variations dans la variance a priori, les résultats que l'on obtiendrait à l'aide d'un estimateur empirique de Bayes ne devraient pas être très différents de ceux indiqués ici pour la version modifiée de cet estimateur.

Àfin d'analyser le biais (au sens classique de l'induction fondée sur un plan), nous avons fait la moyenne des estimations pour les 205 répétitions. La figure 1 donne les moyennes pertinentes pour chacune des 15 sous-régions et chaque méthode d'estimation. Les taux d'activité

de variables explicatives, tant quantitatives que qualitatives, se rattachant à la personne μ_{ij} et $\bar{\Gamma}$ un vecteur des paramètres du modèle. Alors,

$$Z_{\mu_{ij}}^T \bar{\Gamma} = \theta_{\mu} + X_{\mu_{ij}} \beta + \phi_i, \tag{3.16}$$

$$\pi_{\mu_{ij}} = [1 + \exp (-Z_{\mu_{ij}}^T \bar{\Gamma})]^{-1}. \tag{3.17}$$

Par la méthode d'approximation de Taylor, nous pouvons définir la formule de la variance a posteriori de la proportion estimée pour une petite région.

$$\text{Var}(\hat{p}_i) = \left[\sum_{\mu} Z_{\mu_{ij}}^T \pi_{\mu_{ij}} (1 - \pi_{\mu_{ij}}) \right] \frac{\bar{\Sigma}^i}{N_i^2} \left[\sum_{\mu} Z_{\mu_{ij}} \pi_{\mu_{ij}} (1 - \pi_{\mu_{ij}}) \right]. \tag{3.18}$$

Dans l'équation ci-dessous, $\bar{\Sigma}^i$ est la matrice des covariances a posteriori des paramètres estimés de la régression logistique $\bar{\Gamma}$.

Si la taille de l'échantillon dans une petite région représente une partie appréciable de la population de cette région, on pourra accroître le degré de précision en appliquant la méthode de prédiction uniquement aux unités non échantillonnées, comme l'a fait pour la première fois Royall (1970) dans le cas de l'échantillonnage pour population finie.

4. SIMULATION

Nous avons réalisé une simulation pour illustrer les caractéristiques de trois méthodes d'estimation de proportions pour petites régions, à savoir l'estimation non biaisée classique, l'estimation fondée sur un modèle, qui s'apparente à l'estimation "synthétique" de Gonzalez et Hoza (1978), ainsi qu'une version modifiée de l'estimation empirique de Bayes décrite dans la section précédente. La simulation a porté sur un plan de sondage à deux degrés. Les 15 unités primaires d'échantillonnage (UPÉ) représentaient les sous-régions pour lesquelles il fallait estimer un taux d'activité. Un échantillon aléatoire simple de 25 personnes a été prélevé dans chacune des UPÉ, ce qui faisait un échantillon global de 375 personnes. Pour contourner le problème de l'échantillonnage pour population finie, nous avons supposé que les populations des sous-régions étaient infinies.

Comme il s'agissait ici d'évaluer des méthodes d'estimation pour petites régions, nous avons décidé de simuler le ré-échantillonnage au second degré seulement. Autrement dit, nous avons tiré les 15 mêmes UPÉ pour chacune des simulations mais chaque échantillon formé à partir de ces UPÉ était différent. La simulation a consisté en 205 répétitions. Les données ont été générées à l'aide du modèle défini par l'équation (2.3). Les paramètres étaient définis de la façon suivante:

$$\begin{aligned} \theta_1 &= -0.5 \\ \theta_2 &= -1.0 \\ \beta &= 0.1. \end{aligned} \tag{4.1}$$

Les paramètres aléatoires ϕ_i ont été tirés d'une distribution normale de moyenne nulle et d'écart type 0.25. On a pu obtenir les probabilités $\pi_{\mu_{ij}}$ en transformant la fonction logistique comme en (3.15).

$$-\frac{\partial}{\partial^2} \sum_{\mu j} \pi_{\mu j} (1 - \pi_{\mu j}) X_{\mu j} = \frac{\partial \beta}{\partial \phi_i} \frac{\partial \phi_i}{\partial^2}$$

$$-\frac{\partial}{\partial^2} \sum_j \pi_{\mu j} (1 - \pi_{\mu j}) = \frac{\partial \theta_{\mu}}{\partial \phi_i} \frac{\partial \phi_i}{\partial^2}$$

(3.12)

(3.13)

3.2 Estimateurs empiriques de Bayes

Pour déterminer des estimateurs empiriques de Bayes, il faut estimer la variance a priori, σ^2 , à l'aide des données. Nous aurons une estimation fiable de la variance à la condition que l'échantillon renferme un nombre suffisant d'UPB; si le nombre d'UPB est insuffisant, il sera alors préférable d'utiliser une méthode purement bayésienne. Nous nous proposons d'estimer la variance a priori au moyen de l'algorithme EM défini par Dempster, Laird et Rubin (1977). Le modèle général d'estimation est le même que celui utilisé par Laird (1978) pour l'analyse de tableaux de contingence et Tomberlin (1988) pour l'analyse de données de Poisson dans une classification à deux critères. Pour ce qui est des estimations relatives à l'échantillon à deux degrés simple, nous les calculons exactement de la même façon que Leonard (1988).

On déclenche l'algorithme en choisissant une valeur de départ, $\sigma_{(0)}^2$, pour la variance. On détermine ensuite la distribution a posteriori des effets aléatoires, ϕ_i , à l'aide d'une analyse bayésienne comme celle décrite à la Section 2, puis on utilise la distribution ainsi obtenue pour l'étape B de l'algorithme, où on calcule l'espérance de la statistique exhaustive en fonction des données. Enfin, on procède à l'étape M de l'algorithme en calculant simplement la fonction du maximum de vraisemblance des statistiques exhaustives. Pour une description plus détaillée de l'algorithme EM pour les fonctions de densité exponentielles courantes, voir Dempster, Laird et Rubin (1977). On répète ensuite le processus par une analyse bayésienne fondée cette fois sur la nouvelle valeur estimée de la variance, $\sigma_{(1)}^2$. L'algorithme est ainsi exécuté jusqu'à ce qu'il y ait convergence.

3.3 Estimateurs de proportions pour petites régions

Dans les sections 3.1 et 3.2, nous avons présenté les valeurs estimées des paramètres du modèle défini en (2.3-4), ainsi que les variances et les covariances a posteriori correspondantes. Nous allons maintenant nous servir de ces paramètres estimés pour établir, à l'aide d'une méthode de prédiction, des estimations de proportions pour petites régions. À cette fin, nous faisons la moyenne des probabilités individuelles estimées en supposant que la taille de l'échantillon dans chaque petite région est faible par rapport à la taille de la population correspondante:

$$\hat{p}_i = \frac{\sum_{\mu j} \pi_{\mu j}}{N_i}$$

(3.14)

où N_i est le nombre de personnes dans la petite région i et où la probabilité estimée se rattache à la personne μj , $\pi_{\mu j}$, est déterminée par la transformation de la fonction logistique, c.-à-d.

$$\pi_{\mu j} = [1 + \exp\{-(\theta_{\mu} + X_{\mu j} \hat{\beta} + \phi_i)\}]^{-1}.$$

(3.15)

Pour déterminer les variances a posteriori des estimateurs de proportions pour petites régions, il est indiqué de se servir d'une notation plus classique pour la partie linéaire du modèle, notamment en utilisant des variables indicatrices pour désigner les classifications. Soit $Z_{\mu j}$ un vecteur

De cela nous déduisons la distribution a posteriori des paramètres

(3.4)

$$p(\bar{\theta}, \bar{\phi}, \beta \mid y, \sigma^2, X) \propto p(y, \bar{\theta}, \bar{\phi}, \beta \mid \sigma^2, X) \frac{p(y \mid \sigma^2, X)}{p(y \mid \sigma^2, X)}.$$

Il n'est pas possible d'obtenir une expression en forme analytique pour la distribution a posteriori définie ci-dessus à cause de l'intégration très complexe qu'il faudrait exécuter pour obtenir la distribution marginale de y . Dans les circonstances, nous allons recourir à la méthode d'approximation utilisée par Laird (1978) et Tomberlin (1988). Nous exprimons la distribution a posteriori comme une distribution normale multidimensionnelle dont la moyenne coïncide avec le mode de (3.4) et la matrice des covariances égale l'inverse de la matrice d'information calculée au mode.

Pour connaître le mode, nous devons résoudre le système d'équations suivant. Cela peut se faire à l'aide d'un algorithme de Newton-Raphson multidimensionnel.

(3.5)

$$\sum_{mij} y_{mij} X_{mij} = \sum_{mij} \pi_{mij} X_{mij}$$

(3.6)

$$y_{mij} = \sum_{mij} \pi_{mij}$$

(3.7)

$$\sum_{mij} (y_{mij} - \pi_{mij}) - \frac{\sigma^2}{\phi_i} = 0.$$

On détermine la matrice des covariances a posteriori des paramètres en inversant le négatif de la matrice des dérivées secondes du logarithme de (3.4) par rapport aux paramètres, calculée au mode. On notera que le dénominateur de (3.4) ne figure ni dans les équations pour le mode, ni dans la matrice des covariances.

Les éléments de l'inverse de la matrice des covariances a posteriori sont définis comme suit:

(3.8)

$$\frac{-\partial^2}{\partial^2 z} = \sum_{mij} \pi_{mij} (1 - \pi_{mij}) X_{mij}^2$$

(3.9)

$$\frac{-\partial^2}{\partial^2 z} = \sum_{mij} \pi_{mij} (1 - \pi_{mij})$$

(3.10)

$$\frac{-\partial^2}{\partial^2 z} = \sum_{mij} \pi_{mij} (1 - \pi_{mij}) - \frac{\sigma^2}{1}$$

(3.11)

$$\frac{-\partial^2}{\partial^2 z} = \sum_{mij} \pi_{mij} (1 - \pi_{mij}) X_{mij}$$

les effets fixes. Bien qu'il soit facile de formuler des modèles qui correspondent à des plans de sondage à plusieurs degrés sans pousser plus loin la recherche, nous ne pouvons dire encore s'il sera difficile de produire des estimations à l'aide de ces modèles plus complexes.

Dans la réalité, les variables explicatives devraient être définies à l'aide des données. Il faudrait pour cela élaborer une sorte de méthode de sélection de modèle. Bien que l'élaboration de telles méthodes ne soit pas l'objet premier de cet article, il est possible d'imaginer une méthode qui reposerait sur une analyse préliminaire qui ferait usage, par exemple, des méthodes de sélection de variables utilisées normalement pour les modèles de régression logistique, et qui ont été décrites par Haberman (1978). On pourrait réaliser ce genre d'analyse sans tenir compte des paramètres d'effets aléatoires. Après avoir choisi une série de variables explicatives, on introduirait alors les effets aléatoires de la façon prescrite par le plan de sondage.

3. ESTIMATEURS

Dans cette section, nous allons construire des estimateurs empiriques de Bayes pour le modèle simple défini par les équations (2.3-4). Nous supposons premièrement que la variance, σ^2 , est connue et que les estimateurs bayésiens des probabilités π_{ij} sont calculés. Nous nous servons ensuite de l'algorithme EM, décrit par Dempster, Laird et Rubin (1977), pour calculer l'estimateur du maximum de vraisemblance de σ^2 étant donné des estimateurs empiriques de Bayes. Enfin, nous calculons les variances a posteriori de ces estimateurs. La méthodologie est la même que celle exposée dans Laird (1978) et Tumberlin (1988).

3.1 Estimateurs de Bayes

Comme le souligne Laird (1978), dans son analyse de tableaux de contingence, Dempster, Rubin et Tsutakawa (1981), dans leur analyse des composantes de variance pour les modèles linéaires, et Tumberlin (1988), dans son analyse de données de Poisson, on peut faire une analyse bayésienne d'un modèle mixte comme celui défini en (2.3-4) en appliquant une distribution a priori uniforme aux paramètres fixes, θ_μ et β et la distribution a priori définie en (2.4) aux paramètres aléatoires, ϕ_i .

Soit y le vecteur des variables de résultat (0-1) qui indiquent si une personne appartient ou non à la population active et y le vecteur des $\bar{\pi}$, probabilités individuelles π_{ij} . Les données sont alors distribuées selon une binomiale produit définie comme ci-dessous:

(3.1)

$$p(y | \bar{\pi}) \propto \prod_{ij} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{(1 - y_{ij})}.$$

La distribution a priori des paramètres est définie

(3.2)

$$p(\bar{\theta}, \bar{\phi}, \beta | \sigma^2) \propto \exp \left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right].$$

Par conséquent, la distribution conjointe des données, y , et des paramètres est définie

(3.3)

$$p(y, \bar{\theta}, \bar{\phi}, \beta | \sigma^2, X) = p(y | \bar{\theta}, \bar{\phi}, \beta, \sigma^2, X) p(\bar{\theta}, \bar{\phi}, \beta | \sigma^2, X) \propto \left[\prod_{ij} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{(1 - y_{ij})} \right] \exp \left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right].$$

où $\pi^{\mu\nu}$ représente la probabilité d'une "réponse" pour l'unité ν dans la case μ , l'indice μ désigne un ensemble de covariables qualitatives et l'indice ν , un ensemble de caractéristiques d'échantillonnage hiérarchiques (c.-à-d., UPF, USE dans UPF, et ainsi de suite). Le paramètre θ^{μ} représente la somme des effets de classification fixes et le paramètre ϕ^{ν} , la somme des effets aléatoires liés aux caractéristiques d'échantillonnage; le vecteur $X^{\mu\nu}$ est un vecteur de covariables quantitatives et le paramètre β , un vecteur de paramètres de régression linéaire logistique fixes. On suppose que les effets aléatoires ont une distribution paramétrique; il s'agit le plus souvent d'une distribution normale multidimensionnelle. Pour obtenir l'équation de $\pi^{\mu\nu}$, on transforme l'équation (2.1) de la façon suivante,

$$(2.2) \quad \pi^{\mu\nu} = [1 + \exp\{-(\theta^{\mu} + X^{\mu\nu} + \phi^{\nu})\}]^{-1}.$$

Pretons un exemple simple pour illustrer nos propos. Le taux d'activité est la proportion qui nous intéresse en l'occurrence. Supposons que nous avons une variable de classification qui indique le sexe et une covariable continue qui indique l'âge de la personne. Supposons en outre que le plan de sondage prévoit un échantillonnage à deux degrés simple. Le premier degré consiste en un échantillonnage de comtés et le second, en un échantillonnage aléatoire simple de personnes dans chacun des comtés choisis.

$$(2.3) \quad \text{logit}(\pi^{\mu\nu}) = \theta^{\mu} + X^{\mu\nu}\beta + \phi^{\nu}$$

$$(2.4) \quad \phi^{\nu} \sim \text{i.i.d. selon une loi normale}(0, \sigma^2).$$

Dans le modèle ci-dessus, l'indice de classification, μ désigne le sexe de la personne; l'indice des caractéristiques d'échantillonnage, $\nu = ij$, désigne la personne j dans l'UPF i ; $X^{\mu\nu}$ désigne l'âge de la personne et ϕ^{ν} est un effet aléatoire rattaché à l'UPF i .

Le fait de supposer que les effets rattachés aux UPF sont indépendants et identiquement distribués implique que les écarts observés pour les UPF par rapport à la partie fixe du modèle sont interchangeables, c'est-à-dire que, hormis les effets de l'âge et du sexe, il n'existe pas d'information systématique sur les différences de taux d'emploi entre les comtés. Évidemment, ce genre d'information existerait dans une situation réelle, par exemple industrie dominante, distance par rapport aux principaux marchés, ventes au détail, etc. Il faudrait alors incorporer cette information supplémentaire dans le modèle. Cependant, pour les besoins de la cause nous allons conserver notre modèle simple. Pour simplifier la formulation mathématique, nous avons choisi une distribution normale pour les termes d'erreur et il faudra aussi évaluer les conséquences de ce choix après une analyse des données réelles. Il est facile d'ajouter des covariables, qualitatives ou quantitatives, dans le modèle défini en (2.3-4).

En théorie, il est également facile d'étendre le modèle à des plans de sondage plus complexes. Par exemple, on peut modéliser les données recueillies à l'aide d'un plan d'échantillonnage à trois degrés en se servant des effets aléatoires emboîtés, ce qui donne:

$$(2.5) \quad \begin{aligned} \text{logit}(\pi^{\mu\nu}) &= \theta^{\mu} + X^{\mu\nu}\beta + \phi_i + \phi_{j(i)} \\ \phi_i &\sim \text{normale}(0, \sigma_i^2) \\ \phi_{j(i)} &\sim \text{normale}(0, \sigma_j^2). \end{aligned}$$

Dans le modèle ci-dessus, l'indice des caractéristiques d'échantillonnage, $\nu = ijk$ désigne la personne k dans l'USE j , elle-même comprise dans l'UPF i . Le paramètre ϕ_i est l'effet aléatoire rattaché à l'UPF i et $\phi_{j(i)}$ est l'effet aléatoire emboîté rattaché à l'USE j dans l'UPF i . On pourrait aussi inclure les variables de stratification dans la partie du modèle qui représente

et les covariances sont connues et que l'algorithme EM sert dans un deuxième temps à estimer ces paramètres. Non seulement les modèles à effets aléatoires ouvrent la voie à l'estimation par la méthode du maximum de vraisemblance, mais aussi ils permettent de mesurer le degré de fiabilité des estimations finales des paramètres au moyen de variances a posteriori. Ericksen (1980) propose l'utilisation de l'erreur quadratique moyenne (EQM) pour évaluer l'efficacité de la régression dans l'estimation pour petites régions. Il tente de répondre à des questions comme celles-ci: quand doit-on ajouter des variables explicatives dans l'équation de régression? doit-on utiliser les méthodes de pondération de James-Stein lorsque l'estimation synthétique et l'estimation obtenue par régression sont très différentes? une de l'autre? Il précise aussi qu'il ne faut pas négliger l'effet des valeurs aberrantes sur l'estimation calculée et l'erreur estimée correspondante. Peut-être faudrait-il s'intéresser de plus près à l'effet de la non-vérification des hypothèses du modèle linéaire sur les estimateurs pour petites régions. Bien qu'elles aient servi à estimer des totaux pour des statistiques comme celles touchant le chômage et la mortalité, la plupart des méthodes exposées ci-dessus sont conçues surtout pour estimer des variables de résultat continues. Purcell et Kish (1980) ont mis au point une méthode d'analyse de données qualitatives permettant d'estimer des totaux pour petits domaines. Cette méthode consiste essentiellement à ajuster des modèles linéaires logarithmiques aux données, à retrancher quelques-uns des termes d'interaction de degré supérieur et à calculer des estimations au moyen de l'algorithme d'ajustement proportionnel itératif défini par Deming et Stephan (1940). Nous nous proposons d'étendre ces modèles à l'estimation de proportions pour petits domaines (au sens de Dempster et Tomberlin (1980)) en appliquant des méthodes empiriques de Bayes à des modèles de régression logistique avec effets aléatoires. Cela aurait notamment pour avantage de produire une mesure de la variabilité des estimations régionales, à savoir les variances a posteriori approximatives. L'estimateur que nous proposons ici est de même nature que l'estimateur composé dont se servent Schabale et coll. (1977) pour établir les taux de chômage, les deux estimateurs ne différant essentiellement que par la façon de choisir les poids. Nous croyons toutefois que le modèle empirique de Bayes offre une méthode plus naturelle et plus intuitive pour calculer les poids. L'estimation empirique de Bayes fonde sur des effets aléatoires logistiques simples s'est déjà avérée utile dans l'étude de la variation des taux de mortalité selon les régions (voir Miao 1977). Des modèles à effets aléatoires un peu plus complexes ont été utilisés pour estimer des proportions à l'aide de données de l'Enquête mondiale sur la fécondité (Wong et Mason 1985) et des paramètres de Poisson à l'aide de données sur l'assurance-automobile (Weisberg, Tomberlin et Chatterjee 1984 et Tomberlin 1988).

Roberts, Rao et Kumar (1987) ont ajusté des modèles de régression logistique à des données binaires obtenues au moyen de sondages complexes, construit des pseudo-estimateurs du maximum de vraisemblance et comparé leurs estimations à des estimations non biaisées. Ils ont aussi proposé un test de validité de l'ajustement pour leur modèle, qui tient compte du plan de sondage. La différence fondamentale entre notre méthode et celle de Roberts et coll. est qu'en incluant les caractéristiques du plan de sondage dans le modèle, nous pouvons estimer les paramètres et avoir une bonne idée de la fiabilité de ces estimations grâce aux méthodes du maximum de vraisemblance.

2. LE MODÈLE

Conformément au modèle d'analyse de Dempster et Tomberlin (1980) dans sa forme la plus générale, nous définissons un modèle qui exprime les probabilités liées aux unités de la population en fonction de variables qualitatives, de covariables continues et de caractéristiques d'échantillonnage. Les modèles analysés dans cet article sont des cas particuliers du modèle suivant:

$$\text{logit}(\pi_{\mu\nu}) = \theta_{\mu} + X_{\mu\nu}^T \beta + \phi_{\nu} \quad (2.1)$$

un outil intéressant et virtuellement puissant. Cette méthode d'estimation a fait l'objet, ces dernières années, de nombreuses études empiriques portant sur des données réelles aussi bien que des données simulées; notons au passage celles de Levy (1971), Gonzalez (1973), Gonzalez et Hoza (1978) et Schabale (1979). Plusieurs de ces études sont décrites dans un ouvrage colligé par Platek et Singh (1986).

Royall (1970, 1973) s'est aussi intéressé à l'estimation de totaux pour des populations finies lorsqu'il existe de l'information supplémentaire; il se sert pour cela d'une méthode fondée sur un modèle. Après avoir défini un modèle probabiliste de la relation entre la variable d'intérêt et la variable auxiliaire, il détermine des prédicteurs optimaux pour sous-domaines.

Holt, Smith et Tomberlin (1979) et Laake (1979) ont appliqué la méthode de prédiction de Royall à l'estimation pour petites régions. Laake (1979) a observé que, contrairement à la méthode d'estimation synthétique, par laquelle on obtient des estimateurs biaisés sans pour autant disposer d'une méthode explicite pour en estimer le biais, la méthode de prédiction produisait des estimations de l'erreur quadratique moyenne (EQM) dont on pouvait se servir pour comparer les estimateurs. En ce qui concerne l'estimation de totaux pour de petites régions, Holt, Smith et Tomberlin (1979) ont défini diverses formes de structure de population afin de modéliser la relation supposée entre les sous-régions. Une fois que l'on a défini un modèle, il est alors possible de vérifier si les données sont en accord avec ce modèle et aussi d'analyser l'effet d'un vice de modèle sur le biais des estimateurs observés. La variance de l'estimateur, la valeur estimée de cette variance et l'EQM varient selon les modèles. Holt, Smith et Tomberlin (1979) ont construit des intervalles de confiance fondés sur un modèle, que l'on interprète en fonction de réalisations répétées selon le modèle de superpopulation.

Purcell et Kish (1979, 1980) ont examiné les diverses méthodes d'estimation en usage pour les petites régions; ils les ont divisées en cinq grandes catégories: méthodes axées sur la régression, méthodes bayésiennes et méthodes empiriques de Bayes, théorie de la prévision pour superpopulation, méthodes de classification automatique et méthodes d'analyse de données qualitatives. Ils ont fait valoir que l'estimation pour petit domaine ne devait pas être envisagée dans une perspective globale et qu'il y avait beaucoup d'autres facteurs, comme la taille du domaine, dont il fallait tenir compte dans le choix de l'estimateur. Cette affirmation devait être entendue ultérieurement par Särndal (1984).

La grande faiblesse des estimateurs fondés sur un modèle est que les modèles à effets fixes ne permettent pas d'obtenir des estimations de l'erreur quadratique moyenne utiles puisque les valeurs estimées de la variance correspondante ne reflètent pas le biais dont sont entachées inévitablement les estimations fondées sur des modèles auxquels il manque des paramètres. On a donc envisagé l'estimation pour petites régions de deux façons différentes.

Fay et Herriot (1979) ont appliqué la théorie de l'estimation de James-Stein (James et Stein 1961) à des données d'échantillon afin d'établir des estimations du revenu pour de petites régions tirées du recensement de la population et du logement de 1970 aux Etats-Unis. De fait, ils ont utilisé une méthode empirique de Bayes qui avait été mise de l'avant par Robbins (1955) et reprise par Efron et Morris (1975), donnant ainsi une forme précise à l'intéressante proposition de Madow et Hansen (1975), qui avaient suggéré de faire la moyenne pondérée des estimations de l'échantillon et des estimations obtenues par régression. Par un raisonnement semblable, Schabale et coll. (1977) en arrivent à une méthode permettant de calculer un estimateur composé qui est la moyenne pondérée de l'estimateur sans biais et de l'estimateur synthétique. Stroud (1987) et Cressie (1988) donnent d'autres exemples de méthodes empiriques de Bayes fondées sur la théorie normale, qui servent à l'estimation pour petites régions.

Battese, Harter et Fuller (1988) proposent un modèle de régression à erreurs emboîtées pour estimer les moyennes et à cette fin, ils utilisent une méthode de prédiction. Dempster, Rubin et Tsutakawa (1981) avaient proposé antérieurement un modèle plus général pour estimer les moyennes; il s'agissait d'un modèle de régression à coefficients aléatoires. Ces auteurs utilisent des méthodes bayésiennes pour estimer les effets fixes et les effets aléatoires dans des modèles de composantes de la covariance, où l'on suppose, non sans hésitation, que les variances

estimations de proportions pour petites régions à l'aide d'un modèle bayésien. Cette méthode s'écarte sensiblement des méthodes d'estimation synthétique de Gonzalez et Hoza (1976, 1978), de Gonzalez et Wakseberg (1975) et d'autres, qui, elles, sont fondées implicitement sur un modèle.

Comme une enquête complexe sera souvent caractérisée par une structure hiérarchique (c.-à-d. unités primaires d'échantillonnage (UPÉ), unités secondaires d'échantillonnage (USE) dans les UPÉ, unités tertiaires d'échantillonnage (UTE) dans les USE et, finalement, mélanges dans les UTE), la méthode fondée explicitement sur un modèle nous permettra de tenir compte de la complexité du plan de sondage. L'introduction d'un modèle avec effets aléatoires a pour but de trouver, par des méthodes empiriques de Bayes, une solution de compromis entre les estimateurs non biaisés classiques, qui reposent uniquement sur des données propres à des sous-régions, et les estimateurs de modèle à effets fixes, qui reposent sur des données de toutes les régions.

Dans la Section 1.2, nous donnons un compte rendu critique des articles et ouvrages qui ont été rédigés sur la question et proposons une solution au problème de l'estimation de proportions pour petites régions. Le modèle et les estimateurs correspondants sont exposés dans les sections 2 et 3 respectivement. Enfin, dans la dernière section, nous appliquons les résultats de l'analyse à des données simulées tirées d'une étude de Monte Carlo.

1.2 Compte rendu critique et proposition d'une solution

Devant le besoin sans cesse croissant de données régionales et l'impossibilité d'obtenir directement des estimations fiables pour les petites régions ou les sous-domaines par les méthodes de sondage classiques, plusieurs statisticiens se sont intéressés de plus près à l'estimation pour petites régions. Pour cela, ils ont dû recourir à des méthodes fondées implicitement ou explicitement sur un modèle, par lesquelles on va "puiser" dans les petites régions pour accroître la taille effective de l'échantillon à des fins d'estimation et, par conséquent, obtenir des estimations plus précises. Bien que la recherche faite jusqu'à maintenant ait surtout porté sur des modèles linéaires et l'estimation de moyennes ou de totaux, au lieu de proportions, une brève analyse des ouvrages et des articles portant sur ces estimateurs et les critères utilisés pour leur évaluation peut s'avérer très éclairante.

La théorie classique dit qu'un estimateur devrait être convergent selon le plan et, dans la mesure du possible, essentiellement non biaisé selon le plan. Toutefois, ce genre d'estimateur n'est pas toujours très utile lorsque la taille des échantillons est faible.

Gonzalez (1973) décrit la méthode d'estimation synthétique en ces termes: "Une enquête par sondage produit une estimation sans biais pour une grande région; lorsqu'on se sert de cette estimation pour établir des estimations pour les sous-régions en supposant que celles-ci ont les mêmes caractéristiques que la grande région, on dit qu'il s'agit là d'estimation synthétique." (TRADUCTION) Le U.S. National Center for Health Statistics (1968) aurait été le premier à recourir à l'estimation synthétique pour calculer les taux d'invalidité à court et à long terme dans les Etats. Divers auteurs ont tenté par la suite de formaliser la notion d'estimation synthétique, plus particulièrement en ce qui concerne les moyennes de variables de résultat continues, en utilisant des méthodes *appropriées* et des méthodes fondées sur un modèle. Gonzalez (1973), Gonzalez et Wakseberg (1975), Gonzalez et Hoza (1976) de même que Levy et French (1978) ont exécuté une stratification a posteriori à l'aide de données de recensements antérieurs; les strates ainsi formées ont servi à grouper de l'information provenant de toutes les petites régions dans l'hypothèse que la réponse moyenne est la même pour une série de petites régions. Levy (1971), Erickson (1973, 1974) et O'Hare (1976) ont utilisé des méthodes de régression pour introduire de l'information supplémentaire dans l'estimation pour petites régions. La moyenne des erreurs quadratiques moyennes d'échantillonnage pour toutes les sous-régions d'une région donnée est le critère sur lequel on s'est fondé pour évaluer la précision de cette méthode. Erickson (1974) précise que l'existence pas de méthode systématique pour évaluer le biais ou la précision des estimateurs synthétiques. Malgré cette lacune, l'estimation synthétique demeure

Estimation de proportions pour petites régions par des méthodes empiriques de Bayes

BRENDA MacGIBBON¹ et THOMAS J. TOMBERLIN²

RÉSUMÉ

Les auteurs se servent de méthodes empiriques de Bayes pour estimer des proportions pour de "petites régions". Ces méthodes se sont avérées profitables antérieurement dans diverses situations, comme en fait foi l'article de Morris (1983). Suivant l'idée originale de Dempster et Tomberlin (1980), il s'agit essentiellement d'intégrer des effets aléatoires et des effets aléatoires emboîtés dans des modèles qui reflètent la structure complexe d'un plan de sondage à plusieurs degrés. Cela permet d'obtenir des estimations de proportions ainsi que les estimations de variabilité correspondantes. Les auteurs appliquent ces méthodes à des données simulées provenant d'une étude de Monte Carlo qui vise à comparer plusieurs méthodes d'estimation pour petites régions.

MOTS CLÉS: Régression logistique; modèles à effets aléatoires; estimation de Bayes; algorithme EM.

1. INTRODUCTION

1.1 Position du problème

Les enquêtes complexes à plusieurs degrés servent à l'estimation de proportions dans beaucoup de disciplines (par ex.: épidémiologie, économie, criminologie, etc.). Non seulement il est nécessaire d'établir des estimations pour des secteurs restreints et autres sous-groupes particuliers, mais aussi faut-il pouvoir évaluer avec une certaine fiabilité la précision de ces estimations. De là la nécessité de mettre au point de meilleures méthodes d'estimation et d'inférence statistique.

Par surcroît, les méthodes fondées sur la théorie normale et dont se servent Fay et Herriot (1979) pour estimer le revenu (une variable aléatoire continue) dans les petites régions ne se prêtent plus vraiment à l'estimation de proportions pour des variables de résultat discontinues. Dans le présent article, c'est le logit de la proportion, et non la proportion proprement dite, qui fait l'objet d'un modèle linéaire. Cela n'élimine pas les problèmes d'estimation que l'on retrouve dans la théorie classique de la régression logistique (voir Haberman 1978). Malheureusement, en ce qui concerne l'estimation pour petites régions, on s'est encore moins attaché à résoudre ces problèmes, évidemment plus complexes dans les circonstances.

Pour analyser les problèmes liés à l'application des résultats d'une enquête complexe (à plusieurs degrés) relativement restreinte à de petites régions ou de petits domaines qui ne sont pas nécessairement couverts par l'enquête, nous avons choisi une méthode fondée explicitement sur un modèle. Dempster et Tomberlin (1980) l'avaient mise de l'avant pour estimer le taux de sous-dénombrement dans le recensement à l'aide des données d'une enquête postcensitaire. Cette méthode intègre un modèle de régression logistique multiple avec effets aléatoires et des techniques empiriques de Bayes. Elle permet d'estimer directement la variance des

¹ Brenda MacGibbon, Département des sciences de la décision et des systèmes d'informatique de gestion, Université Concordia, 1455 Ouest, boul. de Maisonneuve, Montréal (Québec) H3G 1M8 et Département de mathématiques et d'informatique, Université du Québec à Montréal, C.P. 8888, succ. "A", Montréal (Québec) H3C 3P8.
² Thomas J. Tomberlin, Département des sciences de la décision et des systèmes d'informatique de gestion, Université Concordia, 1455 Ouest, boul. de Maisonneuve, Montréal (Québec) H3G 1M8.

- CHAUDHURI, A., et MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker, Inc..
- CHOW, L.P., LIU, P.T., et MOSELY, W.H. (1973). A new randomized response technique for study of contemporary social problems. Présenté à la 101^{ème} conférence annuelle de l'American Public Health Association, Statistics Section.
- FRANKLIN, L.A. (1977). A Bayesian approach to randomized response sampling. Thèse de doctorat non-publiée, Indiana University, Bloomington, IN.
- GOULD, A.L., SHAH, B.U., et ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Section on Social Statistics American Statistical Association*, 351-359.
- HORVITZ, D.G., GREENBERG, B.G., et ABERNATHY, J.R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistics Review*, 44, 181-196.
- KENNEDY, W.J., et GENTLE, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker, Inc..
- KNUTH, D.E. (1969). *Semi Numerical Algorithms*, (Volume 2). New York: Addison Wesley.
- LIU, P.T., et CHOW, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.
- O'HAGAN, A. (1987). Bayes linear estimates for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.
- PITZ, G.F. (1980). Bayesian analysis of randomized response models. *Psychological Bulletin*, 87, 209-212.
- POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.
- SMOUSE, E.P. (1984). A note on Bayesian least squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- WINKLER, R.L., et FRANKLIN, L.A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.

$|\mu_{g_j} - \mu_{h_j}|$ et réduire $\sigma_{g_j}^2$ et $\sigma_{h_j}^2$ pour $j = 1, 2, 3$. Toutefois, si nous faisons cela, le répondant s'apercevra que, malgré la randomisation, la réponse qu'il donne est susceptible de le trahir et donnera donc peut-être une fausse réponse ou n'en donnera pas du tout. La détermination de valeurs optimales pour les moyennes et les écarts types exige d'autres recherches. Les résultats du tableau 1 donnent un aperçu de l'effet d'une modification de l'écart type. Du point de vue pratique toutefois, les résultats de l'enquête semblaient indiquer que le fait d'avoir choisi des moyennes séparées l'une de l'autre par deux écarts types avait permis de gagner la confiance du répondant et de récupérer (grâce aux essais multiples) 75 % à 85 % de la taille d'échantillon originale sans susciter la "méfiance" que l'on observe si souvent chez les personnes qui participent à des enquêtes à essais multiples.

Plus particulièrement, l'enquête sur le terrain nous a permis de comparer la méthode de l'interview directe avec la méthode des réponses randomisées axées sur l'utilisation d'un appareil électronique; à cette fin, nous avons posé pour les distributions de randomisation normales $\mu_{h_j} = 40$ et $\mu_{g_j} = 50$ de même que $\sigma_{h_j}^2 = \sigma_{g_j}^2 = 5$ pour $j = 1, 2, 3$. Cinq questions délicates ont été posées à deux groupes d'étudiants différents; nous avons observé que la méthode des réponses randomisées produisait des estimations beaucoup plus élevées ($p < .001$) que la méthode de l'interview directe dans trois cas sur cinq. De plus, 88,9 % des personnes qui ont répondu aux questions à l'aide de l'appareil électronique étaient d'avis que leurs amis seraient plus susceptibles de donner des réponses honnêtes s'ils devaient répondre à des questions délicates à l'aide de cet appareil. Il semble donc que cette technique de randomisation a fourni des réponses plus justes que la méthode de l'interview directe (du moins en ce qui concerne certaines questions).

La question de la protection de la vie privée du répondant mérite d'être analysée. Il est contraire à la morale de dire au répondant que la randomisation garantit l'anonymat de ses réponses alors que l'enquête dispose d'un moyen discret pour associer la réponse à la personne (par ex. en utilisant uniquement des nombres pairs pour "OUI" et des nombres impairs pour "NON"). Grâce à l'appareil électronique dont nous avons parlé plus haut, il semble désormais possible de respecter la vie privée des répondants sans perdre trop d'information. Si les moyennes et les écarts types sont mémorisés dans l'appareil et inconnus de l'intervieweur, celui-ci pourra très difficilement identifier les personnes du groupe A et celles du groupe B au cours de l'interview, surtout si l'ordre normal des chiffres qui forment la réponse est modifié. Ainsi, le système qui nous permet d'obtenir plus d'information sans nuire au répondant empêche du même coup l'intervieweur de savoir à quel groupe appartient un répondant.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance aux arbitres et à un rédacteur associé pour leurs remarques constructives.

BIBLIOGRAPHIE

ABERNATHY, J.R., GREENBERG, B.G., et HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.

BARNARD, G.A. (1976). Discussion on the invited and contributed papers. *International Statistical Review*, 44, 226.

BREWER, K.R.W. (1981). Estimating marijuana usage using randomized response some paradoxical findings. *Australian Journal of Statistics*, 23, 139-148.

CAMPBELL, C., et JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27, 229-231.

Premièrement, comme nous l'avons dit dans l'introduction, un dispositif électronique pourrait être plus efficace que les techniques de randomisation classiques, comme les cartes ou la roulette, puisque ces techniques risquent d'être mal appliquées par le répondant ou l'intervieweur, ce qui aurait pour conséquence d'introduire un élément d'erreur incontrôlable. (Voir Abernathy, Greenberg et Horvitz (1970) pour une analyse des problèmes que soulèvent des "cartes mal battues" ou des "cartes perdues" ainsi qu'une formation déficiente des intervieweurs.) Comme la production des valeurs aléatoires est désormais automatique, on n'a plus les problèmes de distribution que pouvait causer l'utilisation des cartes, des boules ou des roulettes puisque la "sélection aléatoire des valeurs" est une opération qui n'est plus exécutée par l'intervieweur et le répondant mais par l'ordinateur. Si l'appareil tombe en panne, ce sera le plus souvent à cause d'une défaillance de la microplaque, incident facilement détectable et très peu probable.

Le second avantage de l'appareil électronique (et peut-être le plus important) est qu'il offre le choix entre deux nombres formés chacun de six chiffres, lequel choix est déterminé par la réponse qu'a l'intention de fournir le répondant ("oui" ou "non"). Mais cette réponse formée de six chiffres renferme en réalité trois réponses constituées chacune de deux chiffres, ce qui représente les trois essais prévus pour chaque répondant. Ainsi, l'expérimentateur profite des avantages des essais multiples sans devoir en subir les inconvénients habituels (mentionnés par Liu et Chow 1976) étant donné que le répondant ignore qu'il se prête à trois essais.

Par surcroît, l'utilisation d'un appareil électronique offre plus de souplesse que les méthodes de randomisation classiques puisque l'expérimentateur peut déterminer à sa guise les six moyennes et les six écarts types. Par exemple, si l'expérimentateur a l'impression que la différence entre les deux premiers chiffres du premier nombre et les deux premiers du second nombre frappent beaucoup les répondants, il peut rapprocher les moyennes μ_{h1} et σ_{h1} et les écarts types μ_{g1} et σ_{g1} , (ou même les faire coïncider). De même, si c'est la différence entre les deux chiffres médians du premier nombre et les deux chiffres médians du second qui attire le moins l'attention, l'expérimentateur peut essayer de tirer le plus d'information possible de ces valeurs en éloignant le plus possible l'une de l'autre les moyennes μ_{h2} et μ_{g2} . On peut aussi intervenir les chiffres. Par exemple, la première valeur aléatoire pourrait être constituée, dans l'ordre, des cinquième et second chiffres au lieu des deux premiers. Cette caractéristique, allée au fait de pouvoir choisir librement les paramètres, devrait fournir une méthode d'échantillonnage qui renseignera beaucoup l'enquêteur sans pour autant nuire au répondant.

Précisons aussi que même si la distribution de randomisation normale était indiquée pour le genre de microprocesseur dont était pourvu l'appareil, nous aurions pu utiliser plusieurs autres distributions continues (ex.: uniforme, Weibull) ou même des distributions discontinues plurivalentes (ex.: multinomiale ou Poisson). La recherche devrait être orientée vers la création de nouveaux microprocesseurs et l'analyse d'autres distributions de randomisation. Par exemple, le coût de production de l'appareil électronique est de l'ordre de 1,500 à 2,000 \$ (puisque il renferme un microprocesseur). En revanche, le coût relativement élevé de cet appareil pourrait être compensé par une forte fréquence d'utilisation et une durée de vie utile appréciable, deux caractéristiques attribuables à la polyvalence de l'appareil.

Il est plus difficile d'évaluer l'attitude du répondant devant cet appareil et la confiance ou la méfiance que celui-ci peut lui inspirer. Les répondants craignent-ils que l'appareil garde en mémoire leurs réponses et que celles-ci soient déchiffrées tard ou tard pour les confondre? Les résultats de l'enquête semblent indiquer que l'on a obtenu un plus grand nombre de réponses honnêtes avec l'appareil électronique qu'avec l'interview direct. Néanmoins, il serait bon de pousser plus loin la comparaison entre cette nouvelle technique de randomisation et les techniques plus classiques.

En pratique, l'étude de questions relatives à la conception de l'appareil (c.-à-d., la détermination des moyennes et des écarts types) renvoie à plusieurs autres considérations. Si nous voulons obtenir plus d'information pour une taille d'échantillon donnée, nous devons accroître

Tableau 2

Valeurs estimées de θ et résultats du test d'hypothèse concernant les valeurs de θ 's obtenues par la méthode de l'interview directe et la méthode des réponses randomisées avec des échantillons de taille $n_1 = 473$ et $n_2 = 477$ respectivement

Question	θ_{1d}	θ_{1r}	n_1^*	valeur z	valeur p
1	.0634	.2013	394.5	6.098	<.0001
2	.1797	.2941	408.1	3.997	<.0001
3	.1078	.1207	384.8	.583	.2810
4	.1882	.1942	409.5	.234	.4091
5	.0042	.0355	339.0	3.341	.0004

Par ailleurs, il est éclairant d'examiner les résultats non significatifs des questions 3 et 4. Ces résultats pourraient amener un observateur à conclure que la méthode des réponses randomisées n'est pas particulièrement supérieure à la méthode de l'interview directe (si l'on fait abstraction des résultats significatifs observés pour les trois autres questions). Cependant, compte tenu justement des trois écarts significatifs observés, on serait peut-être porté à justifier les résultats non significatifs en disant que la question n'était pas "suffisamment délicate" pour qu'un écart appréciable puisse être observé ou même que la question était "si délicate" que le répondant avait préféré mentir même s'il utilisait l'appareil électronique. En outre, la question 1 (Avez-vous déjà triché à un examen que vous avez passé ici à l'université?) semblait relativement "neutre" aux yeux de l'expérimentateur mais on s'est aperçu après coup que le fait de poser cette question après avoir demandé au répondant son numéro de sécurité sociale indisposait les gens beaucoup plus que nous l'aurions cru au départ. Par conséquent, l'incertitude qui existe à propos de l'efficacité de la méthode des réponses randomisées peut être attribuable à des différences d'opinion entre le répondant d'une part et l'intervieweur ou l'expérimentateur d'autre part en ce qui concerne le "degré de sensibilité" d'une question. Ces aspects méritent un examen plus approfondi.

Enfin, 88,9% des personnes qui ont répondu aux questions au moyen de l'appareil électronique (424 sur 477) croient que leurs amis seraient plus susceptibles de répondre honnêtement à des questions délicates s'ils disposaient de cet appareil. Même si on peut penser que des répondants ont surtout voulu faire plaisir à l'intervieweur en répondant à la dernière question par l'affirmative, le très fort pourcentage observé dans ce sens ainsi que les écarts significatifs mentionnés plus haut tendent à démontrer que la méthode des réponses randomisées est bien reçue et jugée conforme aux règles de protection du secret statistique.

7. ANALYSE

Le modèle élaboré dans cet article permet d'appliquer des distributions de randomisation continues et discontinues à l'échantillonnage dans une population dichotomique. Afin d'écarter la randomisation avec distribution normale, nous avons mis au point un appareil électronique programmé. Cet appareil est portatif, peut recevoir en mémoire des moyennes et des écarts types pour les six distributions normales et permet de réaliser simultanément trois essais dont les résultats sont groupés sous la forme d'un nombre à six chiffres. Ce système présente des avantages et des inconvénients par rapport à d'autres techniques fondées sur la méthode des réponses randomisées.

ont été choisies en conformité avec les résultats de la simulation par ordinateur exposée dans la section 4. Pour chacune des deux enquêtes, des étudiants ont été choisis systématiquement parmi ceux qui circulaient sur le campus (un à tous les cinq passants) et ont été interviewés individuellement. Lorsqu'on abordait un étudiant, on lui exposait brièvement le but de l'enquête et on lui demandait s'il voulait y participer. Moins de 10 % de toutes les personnes abordées par les deux équipes d'enquêteurs ont refusé de prêter leur concours. Lorsqu'une personne acceptait de participer à l'enquête, on lui demandait de produire son numéro de sécurité sociale afin de vérifier si elle était réellement inscrite à l'université. À cette fin, on a dû consulter un répertoire d'étudiants fourni par l'administration de l'université et on a exclu de l'enquête tous les répondants qui n'étaient pas réellement inscrits comme étudiants (moins de 5 % des personnes interviewées).

Le fait de demander à une personne son numéro de sécurité sociale l'amène à penser que l'enquêteur fera inévitablement le lien entre elle et les réponses qu'elle donnera. Cette attitude se retrouve dans beaucoup d'enquêtes (par ex. : enquêtes téléphoniques, questionnaires envoyés par la poste, interviews sur place, etc.) et détermine dans une très large mesure l'honnêteté des réponses de la personne. On croit que c'est justement dans ces circonstances que l'expérimentateur peut tirer le plus grand profit de la méthode des réponses randomisées. La taille des échantillons pour la méthode de l'interview direct et la méthode des réponses randomisées était $n_1 = 473$ et $n_2 = 477$ respectivement. Les cinq questions délicates étaient les suivantes :

- Q1 – "Avez-vous déjà triché à un examen que vous avez passé ici à l'université?"
 Q2 – "Seriez-vous tenté de falsifier votre déclaration de revenu?"
 Q3 – "Voleriez-vous votre employeur?"
 Q4 – "Avez-vous fumé de la marijuana dans les 30 derniers jours?"
 Q5 – "Vous êtes-vous déjà livré à un acte homosexuel?"

On jugeait que les cinq questions étaient suffisamment délicates pour faire ressortir de façon non équivoque les avantages de la méthode des réponses randomisées par rapport à la méthode de l'interview directe. De plus, les personnes qui avaient accepté de répondre aux questions par la méthode des réponses randomisées devaient répondre à la fin à une question libellée comme suit : "Croyez-vous que vos amis seraient plus disposés à donner une réponse honnête s'ils devaient répondre à des questions délicates par cette méthode?" Cette question avait pour but de vérifier si la personne interviewée était vraiment convaincue que cette méthode garantissait l'anonymat et était conforme aux règles de protection des renseignements personnels.

Le tableau 2 donne la proportion estimée de répondants (selon la méthode de l'interview directe (θ_{id}) et la méthode des réponses randomisées (θ_r) qui appartiennent au groupe A pour chaque question i ainsi que la valeur estimée de n_i^* (taille effective de l'échantillon) établie suivant la méthode analysée dans la section 3. Ce tableau contient également les valeurs z d'un test d'hypothèse unilatéral $H_0: \theta_{id} - \theta_r = 0$ contre $H_a: \theta_{id} - \theta_r < 0$, ainsi que les valeurs p observées. Comme on s'est servi des tailles d'échantillon n_1 et n_1^* pour ces tests, le résultat est beaucoup plus modéré que si on avait utilisé n_1 et n_2 .

Il convient de noter que les valeurs estimées de θ sont plus élevées dans les cinq cas avec la méthode des réponses randomisées qu'avec la méthode de l'interview directe. En outre, dans le cas des questions 1, 2 et 5, l'écart entre les estimations établi à l'aide des deux méthodes était statistiquement significatif (valeur $p < .001$ pour ces trois questions). Il semble donc que l'on puisse affirmer que la méthode des réponses randomisées avec certaines questions déli-

cates. Il convient aussi de noter que les valeurs choisies pour μ_{gj} , μ_{hj} , σ_{gj} et σ_{hj} , et $k = 3$ nous ont permis d'obtenir une taille effective n_i^* qui représentait généralement 75 % à 85 % de la taille d'échantillon originale n_2 ce qui nous amène à dire que la méthode des réponses randomisées a permis de "récupérer" la majeure partie de l'information.

5. APPAREIL ELECTRONIQUE PORTATIF POUR LA RANDOMISATION

L'échantillonnage fondé sur la randomisation avec distributions normales et essais multiples offre beaucoup de souplesse à l'expérimentateur, celui-ci pouvant choisir à son gré les moyennes et les variances ainsi que le nombre de répondants et le nombre d'essais par répondant. Cet avantage sera toutefois inutile si on ne peut mettre en pratique le plan d'échantillonnage. Le plan d'échantillonnage fondé sur la randomisation de Bernoulli peut être exécuté de diverses façons (par ex. à l'aide de cartes ou de billes de couleur). Or, le modèle que nous élaborons ici prévoit la génération de valeurs aléatoires normales au moyen d'un appareil portatif. On a construit un appareil électronique autour du microprocesseur Intel 8080 pour générer et afficher des valeurs aléatoires normales. On obtient chaque valeur en additionnant 16 nombres aléatoires distribués uniformément et en transformant cette somme de manière à obtenir un écart aléatoire normal ayant la moyenne et l'écart type voulus. En vertu du théorème limite central, les valeurs obtenues devraient être distribuées approximativement selon une loi normale et des tests sérieux indiquent que les valeurs générées par l'appareil suivent effectivement une distribution comme celle des valeurs aléatoires normales. On a préféré cette méthode à d'autres parce qu'elle est facile à programmer par des instructions machine pour ce qui a trait au microprocesseur Intel 8080. Pour avoir plus de détails sur la génération des valeurs aléatoires normales et l'essai de l'appareil, voir Franklin (1977), Kennedy et Gentle (1980) et Knuth (1969).

La version finale de l'appareil a à peu près la taille d'une boîte de cigares et tient facilement dans la main. Il peut être alimenté par un bloc-piles ou un cordon rallonge. Pour l'affichage, les valeurs aléatoires normales ne sont formées que de deux chiffres et l'appareil est conçu de manière à afficher simultanément six nombres de deux chiffres chacun dans des "fenêtres" pouvant contenir chacune six chiffres. Dans une des fenêtres (fenêtre "Oui") apparaissent les valeurs tirées de g_1, g_2 , et g_3 sous la forme d'un nombre à six chiffres. L'autre fenêtre (fenêtre "Non") sert à afficher les valeurs tirées de h_1, h_2 , et h_3 et qui sont groupées sous la forme d'un nombre à six chiffres. Les six moyennes et les six écarts types sont mémorisés dans l'appareil mais on peut les modifier facilement à l'aide d'un petit clavier mobile. Voici comment se déroule l'interview. Tout d'abord, l'intervieweur pose au répondant une question délicate au sujet du groupe A. Le répondant appuie alors sur un bouton pour mettre en activité l'appareil et en un quart de seconde, deux nombres de six chiffres apparaissent dans les fenêtres. Si le répondant appartient au groupe A, il inscrit le nombre paraissant dans la première fenêtre (fenêtre "Oui"); dans le cas contraire, il inscrit le nombre paraissant dans la seconde fenêtre (fenêtre "Non"). Pour convaincre le répondant du "caractère aléatoire" des valeurs observées, l'intervieweur l'invite à appuyer plusieurs fois sur le bouton et à observer les valeurs affichées; ensuite, il lui pose la question délicate. Bien que $k = 3$, il faut préciser que le répondant a l'impression de fournir une seule réponse, en l'occurrence un nombre à six chiffres; en réalité, cette réponse est le résultat de trois essais. Ainsi, l'expérimentateur profite des avantages des essais multiples par répondant sans en subir les inconvénients habituels.

6. RÉSULTATS DE L'ENQUÊTE ET CONCLUSIONS

Deux enquêtes indépendantes ont été menées simultanément sur le campus d'une grande université auprès des étudiants de cette université. L'une des enquêtes consistait à poser cinq questions délicates par interview directe. L'autre enquête consistait à poser les mêmes questions délicates à d'autres personnes mais cette fois, on a utilisé l'appareil électronique randomisées avec randomisation continue; pour cela, on a utilisé l'appareil électronique décrit dans la section précédente. Pour les besoins de l'étude, nous avons posé $k = 3$, $\mu_{g1} = \mu_{g2} = \mu_{g3} = 40$, $\mu_{h1} = \mu_{h2} = \mu_{h3} = 50$ et $\sigma_{g_j} = \sigma_{h_j} = 5$ pour $j = 1, 2, 3$. Ces valeurs

4. ANALYSE DE L'INCIDENCE DES MOYENNES ET DES ÉCARTS TYPES À L'AIDE D'UNE SIMULATION PAR ORDINATEUR

Afin d'analyser l'effet de certains moyennes et de certaines écarts types pour les distributions de randomisation normales de même que l'incidence que peut avoir sur r^* et n^* la valeur de θ et de k (nombre d'essais), nous avons simulé par ordinateur un échantillonnage avec réponses randomisées en formant de façon répétée des échantillons par un processus de Bernoulli avec paramètre θ et k séries de réponses à deux chiffres pour chaque échantillon. Pour la simulation, nous avons posé $\mu_{gj} = 50$, $\mu_{hj} = 40$, et $\sigma_{gj}^2 = \sigma_{hj}^2 = \sigma$ pour $j = 1, \dots, k$. Nous avons considéré deux valeurs de θ (.10 et .25), deux valeurs de σ (6 et 9), trois valeurs de n (50, 200, et 500), et trois valeurs de k (1, 2, et 3). Nous avons choisi ces valeurs car elles permettent de déceler les écarts aléatoires de deux chiffres qui se chevauchent largement dans la distribution; elles peuvent ainsi servir de données-repères lorsque vient le temps de choisir des moyennes et des écarts types pour de vraies enquêtes. Nous avons produit 25 échantillons pour chacune des 36 combinaisons de paramètres. Nous avons déterminé les valeurs de r^* et n^* pour chaque échantillon et avons consigné dans le tableau I la moyenne de n^* pour les 25 échantillons pour chaque combinaison de paramètres.

Les valeurs moyennes de n^* varient considérablement. Suivant le pire scénario $\sigma = 9$, $\theta = .10$, et un seul essai par répondant, n^* n'équivalait qu'à 10 ou à 15 % de n . Par contre, lorsque $\sigma = 6$, $\theta = .25$, et qu'il y a trois essais par répondant, n^* équivalait à environ 75 % de n . Comme prévu, la valeur moyenne de n^* (taille effective de l'échantillon) augmente lorsque n (nombre de répondants) ou k (nombre d'essais par répondant) augmente. Par ailleurs, une diminution de σ ou une augmentation de θ entraîne également une augmentation de n^* .

Nous avons aussi déterminé, pour chaque combinaison de paramètres, la moyenne et la variance de θ pour les 25 échantillons répétés. Les valeurs moyennes de θ sont très près des valeurs correspondantes de θ , (écart inférieur à 5 %) et la variance de θ tend à augmenter lorsque la valeur moyenne de n^* diminue, ce qui tend à confirmer la simulation.

Tableau I
Valeurs moyennes de la taille effective de l'échantillon (n^*) selon diverses tailles d'échantillon (n) et le nombre d'essais par répondant (k)

n	k	$\theta = .10$			$\theta = .25$		
		$\sigma = 6$	$\sigma = 9$	$\sigma = 6$	$\sigma = 6$	$\sigma = 9$	$\sigma = 9$
50	1	16.2	7.0	17.3	9.2	17.8	23.6
	2	27.3	13.1	30.6			
	3	32.6	18.1	38.2			
200	1	58.3	24.8	79.0	41.2	72.9	97.7
	2	103.1	49.6	124.4			
	3	136.6	77.7	151.0			
500	1	148.4	59.6	196.9	103.6	181.2	242.7
	2	261.1	129.3	309.5			
	3	345.8	193.1	375.6			

La quantité d'information que l'on peut obtenir à propos de θ dépend évidemment du choix des moyennes et des écarts types. D'une part, si $\mu_{g_j} = \mu_{h_j}$ et $\sigma_{g_j} = \sigma_{h_j}$ $j = 1, \dots, k$, θ disparaît de la fonction de vraisemblance et \tilde{z} (l'échantillon) ne fournit aucune information sur θ . D'autre part, si $|\mu_{g_j} - \mu_{h_j}| \rightarrow \infty$ pour tout j , σ_{g_j} et σ_{h_j} étant fixes, ou si $\sigma_{g_j} \rightarrow 0$ et $\sigma_{h_j} \rightarrow \infty$ pour tout j , $|\mu_{g_j} - \mu_{h_j}| \neq 0$, étant fixe et différent de 0, nous sommes effectivement en mesure de savoir à quel groupe appartient chaque répondant et dans ces conditions, l'échantillonage ressemble à un échantillonnage de Bernoulli dans θ .

En utilisant une formule d'approximation de $L(\tilde{z} | \theta)$ élaborée par Winkler et Franklin (1979), nous pouvons évaluer plus facilement l'effet de la randomisation et des essais multiples en fonction de moyennes et d'écarts types précis. Autrement dit, il est possible, pour chaque échantillon, de faire une approximation de la fonction de vraisemblance définie en (2.4) à l'aide de la fonction de vraisemblance approximative

$$L^*(r^*, n^* | \theta) = \theta^{r^*} (1 - \theta)^{n^* - r^*}. \quad (3.1)$$

Si nous calculons les dérivées première et seconde du logarithme de la fonction (3.1) et que nous cherchons à déterminer le maximum (θ) et la courbure à ce maximum, nous obtenons:

$$\hat{\theta} = \frac{r^*}{n^*} \quad (3.2)$$

$$\text{et} \quad \left[\frac{\partial^2 \log L^*(r^*, n^* | \theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} = - \frac{\hat{\theta} (1 - \hat{\theta})}{n^*}. \quad (3.3)$$

Ensuite, si nous calculons la dérivée première du logarithme de la fonction de vraisemblance exacte (2.4) et que nous posons le résultat égal à zéro, nous obtenons l'équation qui permettra de calculer l'estimation la plus vraisemblable de θ :

$$\sum_n \frac{\gamma_i + (1 - \theta)\eta_i}{\gamma_i - \eta_i} = 0 \text{ où } \gamma_i = \prod_k g_j(z_{ij}), \eta_i = \prod_k h_j(z_{ij}). \quad (3.4)$$

On obtient la solution ($\hat{\theta}_j$) de (3.4) au moyen d'une recherche par case. En calculant la dérivée seconde du logarithme de la fonction de vraisemblance exacte (2.4), nous obtenons:

$$\left[\frac{\partial^2 \log L(\tilde{z} | \theta)}{\partial \theta^2} \right] = - \frac{\sum_n \frac{[\gamma_i + (1 - \theta)\eta_i]^2}{[\gamma_i - \eta_i]^2}}{2}. \quad (3.5)$$

En appliquant $\hat{\theta}_j$ dans l'équation (3.5), nous obtenons la courbure de la fonction de vraisemblance logarithmique à $\hat{\theta}_j$ (le maximum). Les équations (3.2) et (3.3) sont deux équations avec deux inconnues, r^* et n^* . En posant (3.2) $= \hat{\theta}_j$ et (3.3) $= \hat{\theta}_j$, et la courbure à ce maximum soient valeurs de r^* et de n^* de manière que le maximum $\hat{\theta} = \hat{\theta}_j$, et la courbure à ce maximum soient les mêmes pour la fonction de vraisemblance logarithmique approximative et la fonction de vraisemblance logarithmique réelle. Par conséquent, on peut appeler l'échantillon avec réponses randomisées \tilde{z} à un échantillon avec réponses non randomisées un échantillon de Bernoulli, dont r^* éléments sur n^* appartiennent au groupe auquel s'intéresse l'expérimentateur. Dans ce sens, on peut considérer n^* comme une mesure approximative de la quantité d'information contenue dans l'échantillon avec réponses randomisées de taille n .

Il s'agit d'estimer θ en se fondant sur les kn observations d'échantillon z_{ij} , $i = 1, \dots, n$ et $j = 1, \dots, k$. Pour des raisons de commodité, nous supposons dans le reste de cet article que G_{ij} et H_{ij} sont parfaitement continues, avec g_{ij} et h_{ij} comme fonctions de densité respectives; le même raisonnement s'applique à la distribution discontinue. La fonction de densité conditionnelle de z_{ij} étant donné θ est $\theta g_{ij}(z_{ij}) + (1 - \theta) h_{ij}(z_{ij})$, et la fonction de vraisemblance pour toute l'expérience est:

(2.1)
$$L(\tilde{z} | \theta) = \prod_n \left[\theta \prod_k g_{ij}(z_{ij}) + (1 - \theta) \prod_k h_{ij}(z_{ij}) \right] \text{ pour } 0 \leq \theta \leq 1,$$

où $\tilde{z} = (\tilde{z}_1, \dots, \tilde{z}_n)$ et $\tilde{z}_i = (z_{i1}, \dots, z_{ik})$.
À l'aide du théorème du binôme, nous pouvons réécrire la fonction de vraisemblance de la façon suivante

(2.2)
$$L(\tilde{z} | \theta) = \sum_n \alpha_i \theta^i (1 - \theta)^{n-i} \text{ où } 0 \leq \theta \leq 1 \text{ et}$$

(2.3)
$$\alpha_i = \sum_c \left[\prod_k \prod_{i \in C_{is}} g_{ij}(z_{ij}) \right] \left[\prod_k \prod_{i \notin C_{is}} h_{ij}(z_{ij}) \right],$$

où C_{i1}, \dots, C_{ic} représentent les $c = \binom{n}{i}$ combinaisons de i éléments parmi n éléments. Dans l'équation (2.2), $\theta^i (1 - \theta)^{n-i}$ est la fonction de vraisemblance de Bernoulli, à condition qu'exactly i répondants appartiennent au groupe A, et α_i est la fonction de vraisemblance de \tilde{z} étant donné i dans l'échantillon.
Nous obtenons une forme particulière de (2.1) lorsque nous supposons que les mêmes distributions de randomisation sont utilisées pour les n répondants. Dans ce cas, $g_{ij} = g_j$ et $h_{ij} = h_j$ pour $i = 1, \dots, n$ et l'équation (2.1) se ramène à

(2.4)
$$L(\tilde{z} | \theta) = \prod_n \left[\theta \prod_k g_i(z_{ij}) + (1 - \theta) \prod_k h_j(z_{ij}) \right] \text{ pour } 0 \leq \theta \leq 1.$$

Quelle que soit la forme de l'équation, il faut effectuer une recherche par case pour déterminer les estimations les plus vraisemblables. Cela est réalisable, en l'occurrence, puisque θ est une variable unidimensionnelle qui ne peut prendre que les valeurs de 0 à 1. Cette opération peut être exécutée facilement à l'aide de techniques de recherche courantes appliquées au logarithme de la fonction de vraisemblance. (Voir, par exemple, Kennedy et Gentle 1980).

3. RANDOMISATION AVEC DISTRIBUTIONS NORMALES

Bien que n'importe quelle distribution continue (*ex. Weibull, uniforme, etc.*) puisse servir de distribution de randomisation dans le modèle exposé ci-dessus, nous allons nous concentrer ici sur la distribution normale. Nous allons, de plus supposer que les mêmes distributions de randomisation sont utilisées pour tous les répondants de sorte que nous aurons la fonction de vraisemblance (2.4). En conséquence, g_j et h_j sont des fonctions de densité normales ayant pour moyennes μ_{gj} et μ_{hj} pour écarts types σ_{gj} et σ_{hj} , respectivement. On peut alors établir un rapport entre ces fonctions de densité normales et la fonction de vraisemblance définie dans la section 2.

modèle à essais multiples était plus susceptible d'éveiller la méfiance du répondant et de l'inclure, par conséquent, à ne pas fournir une réponse honnête. L'article de Horvitz, Greenberg et Abernathy (1976) expose plusieurs autres plans qui prévoient des mécanismes de randomisation discontinue. En outre, l'ouvrage récent de Chaudhuri et Mukerjee (1988) intitulé *Randomized Response: Theory and Techniques* contient un exposé théorique détaillé ainsi qu'une analyse de résultats. Warner (1971) présente un modèle plus général, à randomisation continue ou discontinue, et Pitz (1980), Smouse (1984) et O'Hagen (1987) en font l'analyse dans une perspective bayésienne. Quelques enquêtes ont déjà été réalisées à ce jour; certaines d'entre elles ont montré la supériorité de la méthode des réponses randomisées par rapport aux méthodes d'enquête classiques (par ex.: Gold et coll., 1969 et Liu et Chow 1976) et quelques autres ont produit des résultats incertains (par ex.: Brewer 1981). Cependant, seul Poole (1974) a imaginé une distribution de randomisation continue particulière (uniforme) pour estimer une distribution continue en demandant à des enquêtés d'inscrire une réponse qui avait été préalablement multipliée par un nombre choisi aléatoirement dans une table de nombres aléatoires.

Dans cet article, nous allons considérer un modèle de réponses randomisées pour population dichotomique mais avec distribution de randomisation continue. Selon le modèle original de Warner, la question à laquelle doit répondre l'enquêté est déterminée par le dispositif de randomisation alors que dans le cas de la méthode exposée ici, la question est déterminée par le fait que le répondant appartient ou non au groupe auquel on s'intéresse. Selon cette méthode, des valeurs sont tirées aléatoirement de deux distributions ("oui" et "autre pour "non") et le répondant inscrit la valeur qui correspond à sa situation. En donnant une réponse constituée de plusieurs caractères numériques, le répondant se prête simultanément à plusieurs essais. Cela représente un avantage par rapport aux méthodes habituelles en ce que le répondant croit qu'il a donné une seule réponse alors qu'il s'est soumis, à son insu, à plusieurs essais. La section 2 sert à présenter le modèle général, selon lequel la randomisation peut s'effectuer par n importe quel type de distribution. Dans la section 3, nous étudions le cas particulier de la randomisation selon une distribution normale et analysons une méthode d'approximation visant à évaluer l'incidence de la randomisation et des essais multiples par répondant. Dans la section 4, nous cherchons à évaluer, au moyen d'une simulation par ordinateur, comment le choix des moyennes et des écarts types influence sur l'efficacité des enquêtes où l'on a recours à la randomisation par distribution normale avec essais multiples. La section suivante montre comment mettre en application une distribution de randomisation normale par l'utilisation d'un appareil électronique informatisé qui génère et affiche des valeurs aléatoires normales. On croit que cet appareil pourrait s'avérer plus efficace que les cartes ou la roulette puisque ces "outils" peuvent être mal utilisés par le répondant ou l'intervieweur. Dans la section 6, nous analysons les résultats d'une enquête où l'appareil en question a servi à répondre à cinq questions délicates. Enfin, la section 7 contient un résumé de l'étude et une brève analyse de questions relatives au plan de sondage.

2. PRÉSENTATION DU MODÈLE

Supposons que nous cherchons à connaître θ , la proportion d'individus qui appartiennent au groupe A dans une population particulière. Nous tirons un échantillon aléatoire simple de n personnes dans la population $n \geq 1$, en supposant que n est suffisamment faible par rapport à la population pour que l'on puisse envisager un échantillonnage avec remise. Chaque répondant est soumis à k essais $k \geq 1$. À l'essai j pour le répondant i , des valeurs aléatoires sont tirées des fonctions de distribution G_{ij} et H_{ij} . Le répondant voit ces deux valeurs et doit inscrire celle provenant de G_{ij} s'il appartient au groupe A et celle provenant de H_{ij} dans le cas contraire. L'enquêteur connaît la forme exacte de G_{ij} et H_{ij} mais ne voit que la valeur inscrite par le répondant, qui est désignée par z_{ij} , et ne sait donc pas de quelle distribution elle origine.

Echantillonnage pour populations dichotomiques par la méthode des réponses randomisées avec randomisation continue

LEROY A. FRANKLIN¹

RÉSUMÉ

L'auteur élabore un modèle de randomisation des réponses pour des populations dichotomiques. Ce modèle prévoit l'utilisation de la randomisation continue ainsi que des essais multiples pour chaque répondant. L'auteur se penche sur le cas particulier de la randomisation avec des distributions normales et exécute une simulation par ordinateur pour découvrir les effets que peut avoir cette méthode d'échantillonnage sur la quantité d'information dans l'échantillon. Il décrit aussi un appareil électronique portatif qui mettrait en application son modèle. Enfin, il présente les résultats de l'enquête qu'il a réalisée à l'aide de cet appareil. Les résultats illustrent la supériorité de la méthode des réponses randomisées par rapport à l'interview directe, du moins en ce qui a trait à certaines questions délicates.

MOTS CLÉS: Réponses randomisées; randomisation avec distributions continues; simulation par ordinateur.

1. INTRODUCTION

Les enquêtes ont souvent pour but d'estimer la proportion d'individus qui remplissent une condition particulière. Si cette condition a trait à quelque chose de très personnel ou controversé (par ex., recherche d'un nouvel emploi, comportement sexuel) ou à quelque chose d'illégal (par ex., consommation de drogues, activités criminelles), le répondant hésitera peut-être à donner une réponse franche ou refusera peut-être de répondre à une question qui lui serait posée directement sur le sujet. Ainsi, l'estimation de proportions devient problématique lorsqu'on doit se fonder sur une enquête où des questions délicates sont posées directement aux répondants.

Les plans d'échantillonnage fondés sur la méthode des réponses randomisées prévoient un dispositif aléatoire (ou dispositif de randomisation) qui permet à des personnes de répondre à des questions délicates sans se trahir. Le résultat du dispositif est connu du répondant mais non de l'intervieweur. En revanche, l'expérimentateur connaît les propriétés du dispositif et peut, par conséquent, tirer des conclusions sur la proportion à estimer sans savoir quoi que ce soit des personnes qui ont répondu au questionnaire. Le dispositif aléatoire introduit du bruit dans le processus de collecte des données mais il peut être préférable de subir une perte d'information que de devoir composer avec le bruit incontrôlable engendré par la non-réponse ou le mensonge lorsque les questions sont posées directement.

Le modèle des réponses randomisées a été initialement imaginé par Warner (1965); il prévoyait alors une randomisation dichotomique pour une population dichotomique. Winkler et Franklin (1979) ont étudié ce modèle dans une perspective bayésienne. Gould, Shah et Abernathy (1969) ont imaginé le modèle des réponses randomisées avec deux essais ou plus par répondant et Liu et Chow (1976) lui ont apporté des perfectionnements. Dans les deux cas, les auteurs ont prouvé que le modèle à essais multiples produisait des estimations plus efficaces que le modèle à essai unique de Warner. En revanche, ils ont fait remarquer que le

¹ LeRoy A. Franklin, Department of System and Decision Sciences, Indiana State University, School of Business, Terre Haute, Indiana 47809.

- BULL, S.B., et PEDERSON, L.L. (1987). Variance for polychotomous logistic regression using complex survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
 CHAMBLESS, L.E., et BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, Theory and Methods*, 14, 1377-1392.
- COX, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.
- DALE, J.R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society, Sér. B*, 48, 48-59.
- FAY, R.E. (1985). A jackknife chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., et PARK, H.J. (1986). *PC CARP*. Statistical Laboratory, Iowa State University, Ames, Iowa.
- GALLANT, A.R. (1987). *Nonlinear Statistical Methods*. New York: John Wiley & Sons.
- HABERMAN, S.J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.
- HOLT, D., SCOTT, A.J., et EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Sér. A*, 143, 303-320.
- JENNIRICH, R.I., et MOORE, R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Section on Statistical Computing, American Statistical Association*.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- MOORE, D.S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72, 131-137.
- MOREL, J. (1987). Multivariate nonlinear models for vectors of proportions: A generalized least squares approach. Thèse de doctorat non publiée. Iowa State University, Ames, Iowa.
- NELDER, J.A., et WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Sér. A*, 135, 370-384.
- RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., et FULLER, W.A. (1988). Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes. *Techniques d'enquête*, 14, 63-73.

5. RÉSUMÉ

Nous venons d'exposer une méthode qui permet d'obtenir des estimateurs normaux asymptotiques des paramètres d'une fonction logistique généralisée avec variable de réponse multinomiale pour plans de sondage complexes. Nous avons vu que l'équation (2.10) représente un estimateur convergent de la matrice des covariances asymptotique selon un plan de sondage complexe; cette équation vient du développement de Taylor. Dans le cas de grands échantillons, la matrice des covariances asymptotique produit des erreurs de première espèce acceptables pour les tests F qui portent sur les paramètres du modèle. Chose plus importante, nous avons montré que la formule de correction définie en (2.13) et (2.14) donne une matrice des covariances qui réduit le biais dû aux petits échantillons. Cette matrice des covariances redressée présente quelques caractéristiques intéressantes:

1. Elle ramène à un niveau plus acceptable l'erreur de première espèce anormalement élevée qui se produit lorsqu'on ne tient pas compte du plan de sondage complexe, et ce plus rapidement que ne le fait la méthode delta habituelle.
2. Elle est définie positive lorsque $H_n(\hat{\beta}_{PSUDO})$ l'est, peu importe que (2.8) soit ou non singulière.
3. Elle est asymptotiquement équivalente à (2.10).

Nous avons exposé les résultats d'une étude de Monte Carlo dans la section 3. Des données satisfaisant la moyenne conditionnelle logistique (2.1) ont été produites suivant deux plans d'échantillonnage en grappes à un seul degré. Cette simulation nous a permis d'analyser notamment l'incidence de la corrélation interne et de l'effet du plan sur le biais relatif des erreurs de première espèce estimées pour les tests F appliqués à $H_0: \hat{\beta} = \beta^0$. Elle a permis de constater que la méthode élémentaire du maximum de vraisemblance suscitait, comme prévu, un biais relatif élevé. En ce qui concerne les petits échantillons, les résultats de la simulation donnent à penser que la matrice des covariances redressée doit être préférée à celle obtenue par la méthode delta habituelle.

REMERCIEMENTS

Cette étude a été amorcée lorsque l'auteur était étudiant à l'Université Iowa State. L'auteur tient à remercier Wayne A. Fuller pour l'avoir initié à ce sujet et lui avoir proposé quelques unes des modifications (relatives aux petits échantillons) qui ont été apportées à la méthode d'estimation.

BIBLIOGRAPHIE

ALBERT, A., et LESAFFRE, E. (1986). Multiple group logistic discrimination. *Computers and Mathematics with Applications*, 12A, 209-224.

BEDRICK, E.J. (1983). Adjusted chi-square tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.

BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A., GRATTON, M.A., HIDIROGLOU, M.A., KUMAR, S., et RAO, J.N.K. (1984). Analyse de données qualitatives d'enquêtes complexes: quelques expériences canadiennes. *Techniques d'enquête*, 10, 155-170.

4. APPLICATION À L'ÉCHANTILLONNAGE STRATIFIÉ ET À DES PLANS DE SONDAGE PLUS COMPLEXES

Il est possible d'étendre la méthode CPLX à l'échantillonnage stratifié en procédant comme suit. Supposons que la population est divisée en $i = 1, 2, \dots, L$ strates. Soit m_{ij} la taille de la grappe j dans la strate i , n_i le nombre de grappes prélevées dans la strate i et $y_{ij\ell}^*$ la réponse multinomiale du ℓ -ième élément de la grappe j dans la strate i , $\ell = 1, 2, \dots, m_{ij}$, $j = 1, 2, \dots, n_i$, $i = 1, 2, \dots, L$. On suppose que $\tilde{\pi}_{ij\ell}^*$, l'espérance de $y_{ij\ell}^*$, satisfait la fonction logist-que (2.1) pour un vecteur explicatif $\mathbf{x}_{ij\ell}^*$ donné.

Il est possible de déterminer un estimateur convergent de $\tilde{\theta}^0$, par exemple $\hat{\theta}^{\text{PSEUDO}}$, en maximisant la fonction

(4.1)
$$L_n(\tilde{\theta}) = \sum_{i=1}^L \sum_{n_i} \sum_{m_{ij}} w_{ij} (\log \tilde{\pi}_{ij\ell}^*)' y_{ij\ell}^*.$$

On exécute l'algorithme (2.5) avec trois indices i, j, ℓ . On applique la formule de correction définie en (2.13) et (2.14) avec

(4.2)
$$n = \sum_{i=1}^L n_i,$$

(4.3)
$$H_n(\hat{\theta}^{\text{PSEUDO}}) = \sum_{i=1}^L \sum_{n_i} \sum_{m_{ij}} w_{ij} \Delta(\tilde{\pi}_{ij\ell}^*) \otimes \mathbf{x}_{ij\ell}^* \mathbf{x}_{ij\ell}^*,$$

(4.4)
$$\mathcal{G} = [(n^* - k)^{-1} (n^* - 1)] \left[\sum_{i=1}^L (n_i - 1)^{-1} n_i (1 - f_i) \sum_{j=1}^{n_i} (\hat{d}_{ij} - \hat{d}_i)(\hat{d}_{ij} - \hat{d}_i)' \right],$$

(4.5)
$$\hat{d}_{ij} = \sum_{m_{ij}} w_{ij} (y_{ij\ell}^* - \tilde{\pi}_{ij\ell}^*) \otimes \mathbf{x}_{ij\ell}^*,$$

(4.6)
$$\hat{d}_i = n_i^{-1} \sum_{j=1}^{n_i} \hat{d}_{ij},$$

(4.7)
$$f_i = \text{taux de sondage de la strate } i, \text{ et}$$

(4.8)
$$n^* = \sum_{i=1}^L \sum_{n_i} m_{ij}.$$

On peut étendre progressivement la méthode d'estimation aux plans d'échantillonnage à plusieurs degrés en maximisant (4.1) à chaque étape jusqu'aux unités du degré précédent. La sommation définie en (4.3) devrait être élargie de manière à inclure toutes les unités d'échantillonnage finales. L'équation fondamentale ici est (4.4). Sa construction doit reposer sur le plan de sondage complexe. Cette opération peut s'avérer difficile dans le cas de l'échantillonnage à plusieurs degrés. Fuller et coll. (1986, p. 82) présentent des résultats pour le cas de l'échantillonnage à deux degrés stratifié.

Tableau 3.8
Biais relatif des 5^e et 95^e percentiles estimés de la statistique “t” calculée pour les coefficients estimés, suivant le plan d’échantillonnage II

Méthode					
n	ζ^2	$\phi^{0.5} - 1$	5 ^e percentile EMV	95 ^e percentile EMV	95 ^e percentile CPLX
20	0.0	0.00	0.01	0.00	0.18
20	0.2	0.41	0.37	0.32	0.09
20	0.4	0.73	0.63	0.57	0.05
20	0.6	1.00	0.79	0.74	0.05
30	0.0	0.00	0.02	0.00	0.16
30	0.2	0.41	0.39	0.38	0.10
30	0.4	0.73	0.68	0.63	0.08
30	0.6	1.00	0.91	0.86	0.07
40	0.0	0.00	0.01	0.00	0.15
40	0.2	0.41	0.39	0.40	0.06
40	0.4	0.73	0.65	0.60	0.09
40	0.6	1.00	0.99	0.89	0.05
50	0.0	0.00	0.01	0.01	0.10
50	0.2	0.41	0.39	0.40	0.04
50	0.4	0.73	0.73	0.72	0.01
50	0.6	1.00	1.00	0.95	0.01
100	0.0	0.00	0.01	0.01	0.05
100	0.2	0.41	0.40	0.37	0.02
100	0.4	0.73	0.72	0.73	0.00
100	0.6	1.00	1.00	1.02	0.01
200	0.0	0.00	0.02	0.01	0.01
200	0.2	0.41	0.40	0.45	0.02
200	0.4	0.73	0.71	0.68	0.01
200	0.6	1.00	1.03	0.95	0.02

Le tableau 3.8 donne le biais relatif (3.1.16) des 5^e et 95^e percentiles de la statistique “t” (3.1.15) pour les méthodes EMV et CPLX. Comme prévu, la méthode EMV donne un biais relatif presque nul lorsqu’il n’y a pas de corrélation interne. Toutefois, ce biais augmente avec ζ^2 . Quant à la méthode CPLX, elle donne, règle générale, un biais relatif peu élevé et même négligeable dans le cas de grands échantillons.

Tableau 3.7
Caractéristiques de ϕ suivant le plan d'échantillonnage II de l'étude de Monte Carlo

Méthode	CPLX		TAYLOR	
	Biais rel.	E.T.	Biais rel.	E.T.

20	0.0	0.48	0.22	0.04
20	0.2	0.16	0.53	0.26
20	0.4	0.05	0.87	0.34
20	0.6	0.01	1.24	0.39
30	0.0	0.49	0.18	0.02
30	0.2	0.25	0.48	0.19
30	0.4	0.19	0.84	0.24
30	0.6	0.16	1.12	0.27
40	0.0	0.38	0.16	0.02
40	0.2	0.22	0.45	0.14
40	0.4	0.16	0.70	0.20
40	0.6	0.16	0.98	0.19
50	0.0	0.27	0.14	0.02
50	0.2	0.15	0.42	0.12
50	0.4	0.12	0.67	0.15
50	0.6	0.11	0.89	0.16
100	0.0	0.12	0.10	0.01
100	0.2	0.06	0.32	0.07
100	0.4	0.05	0.50	0.07
100	0.6	0.06	0.59	0.07
200	0.0	0.05	0.07	0.01
200	0.2	0.03	0.24	0.03
200	0.4	0.02	0.34	0.04
200	0.6	0.02	0.40	0.03

Le tableau 3.7 présente les caractéristiques (selon l'étude de Monte Carlo) de l'estimateur de l'effet du plan défini en (3.1.8) pour les méthodes CPLX et TAYLOR. La seconde semble légèrement supérieure à la première dans le cas des petits échantillons. De façon générale, les deux méthodes donnent des résultats acceptables. Elles semblent s'équivaloir pour ce qui est des grands échantillons.

Tableau 3.6
Caractéristiques du critère chi carré de $H_0: \hat{\beta} = \hat{\beta}^0$ suivant le plan
d'échantillonnage II de l'étude de Monte Carlo

Méthode	EMV			CPLX			TAYLOR		
	Moyenne	Variance	n	Moyenne	Variance	n	Moyenne	Variance	n

20	11.3	18.9	10.2	19.7	40.5	15x10 ²	20	0.0	1	20	11.3	20.3	62.8	106.4	152.6	18.9	10.2	19.7	40.5	15x10 ²
20	28.3	106.4	10.5	18.4	111.3	42x10 ⁵	20	0.4	3	20	28.3	20.3	62.8	106.4	152.6	18.9	10.5	21.4	39.2	11x10 ²
20	35.2	152.6	10.3	18.2	11x10 ³	50x10 ⁹	20	0.6	4	20	35.2	20.3	62.8	106.4	152.6	18.9	10.3	18.2	11x10 ³	50x10 ⁹
30	11.6	21.6	9.4	16.3	22.0	147.3	30	0.0	1	30	11.6	21.8	75.2	117.6	191.0	21.6	9.4	16.3	22.0	147.3
30	21.8	75.2	9.9	17.5	22.7	161.2	30	0.2	2	30	21.8	75.2	75.2	117.6	191.0	21.6	9.9	17.5	22.7	161.2
30	30.4	117.6	9.8	16.5	24.3	224.6	30	0.4	3	30	30.4	75.2	75.2	117.6	191.0	21.6	9.8	16.5	24.3	224.6
30	39.3	191.0	9.5	14.5	24x10 ²	60x10 ⁸	30	0.6	4	30	39.3	75.2	75.2	117.6	191.0	21.6	9.5	14.5	24x10 ²	60x10 ⁸
40	11.6	21.3	9.9	19.4	18.1	86.7	40	0.0	1	40	11.6	22.4	76.5	153.2	223.1	21.3	9.9	19.4	18.1	86.7
40	22.4	76.5	10.4	18.3	18.9	80.8	40	0.2	2	40	22.4	76.5	76.5	153.2	223.1	21.3	10.4	18.3	18.9	80.8
40	31.8	153.2	10.2	17.8	19.2	90.4	40	0.4	3	40	31.8	76.5	76.5	153.2	223.1	21.3	10.2	17.8	19.2	90.4
40	41.4	223.1	10.1	16.9	19.3	104.4	40	0.6	4	40	41.4	76.5	76.5	153.2	223.1	21.3	10.1	16.9	19.3	104.4
50	11.5	19.9	10.6	20.0	16.1	56.9	50	0.0	1	50	11.5	22.7	80.6	160.1	262.3	19.9	10.6	20.0	16.1	56.9
50	22.7	80.6	11.4	23.9	17.5	70.9	50	0.2	2	50	22.7	80.6	80.6	160.1	262.3	19.9	11.4	23.9	17.5	70.9
50	32.3	160.1	11.1	22.9	17.4	73.7	50	0.4	3	50	32.3	80.6	80.6	160.1	262.3	19.9	11.1	22.9	17.4	73.7
50	41.7	262.3	10.7	19.7	17.0	63.8	50	0.6	4	50	41.7	80.6	80.6	160.1	262.3	19.9	10.7	19.7	17.0	63.8
100	11.8	21.5	11.8	25.2	13.9	36.2	100	0.0	1	100	11.8	22.9	87.3	191.8	297.7	21.5	11.8	25.2	13.9	36.2
100	22.9	87.3	11.9	27.0	14.0	38.5	100	0.2	2	100	22.9	87.3	87.3	191.8	297.7	21.5	11.9	27.0	14.0	38.5
100	34.7	191.8	12.3	27.9	14.4	40.7	100	0.4	3	100	34.7	87.3	87.3	191.8	297.7	21.5	12.3	27.9	14.4	40.7
100	45.1	297.7	12.0	25.0	14.1	37.2	100	0.6	4	100	45.1	87.3	87.3	191.8	297.7	21.5	12.0	25.0	14.1	37.2
200	12.0	23.8	12.1	26.3	13.0	30.3	200	0.0	1	200	12.0	24.0	88.6	175.2	320.0	23.8	12.1	26.3	13.0	30.3
200	24.0	88.6	12.4	25.9	13.3	30.0	200	0.2	2	200	24.0	88.6	88.6	175.2	320.0	23.8	12.4	25.9	13.3	30.0
200	34.5	175.2	12.0	23.3	12.8	27.0	200	0.4	3	200	34.5	88.6	88.6	175.2	320.0	23.8	12.0	23.3	12.8	27.0
200	46.8	320.0	12.2	24.0	13.0	27.9	200	0.6	4	200	46.8	88.6	88.6	175.2	320.0	23.8	12.2	24.0	13.0	27.9

Le tableau 3.6 présente les caractéristiques (selon l'étude de Monte Carlo) du critère chi carré de $H_0: \hat{\beta} = \hat{\beta}^0$ (chi carré = 12 x F) pour les trois méthodes d'estimation étudiées. Avec la méthode CPLX, on observe des moyennes qui n'atteignent pas 12 et des variances qui n'atteignent pas 24 lorsqu'il s'agit de petits échantillons. Cette sous-estimation n'existe plus avec de grands échantillons. En ce qui concerne la méthode TAYLOR, on observe des moyennes et des variances excessives pour les petits échantillons. Par exemple, pour $\zeta^2 = 0.6$, la variance est de l'ordre du milliard lorsque n est plus petit ou égal à 30. Pour de grands échantillons, les méthodes CPLX et TAYLOR semblent donner des résultats similaires. La méthode EMV n'est satisfaisante que lorsque $\zeta^2 = 0.00$. Autrement, les moyennes et les variances estimées sont trop élevées.

À partir des six équations ci-dessus, on a produit 1000 séries d'échantillons avec n grappes de taille $m_j = m = 6$, selon les équations (3.2.1) et (3.2.2) pour diverses valeurs de n , ζ^2 et de ϕ . Le tableau 3.5 donne le biais relatif (3.1.14) des erreurs de première espèce estimées résultant de la comparaison des tests F de $H_0: \hat{\beta} = \hat{\beta}^0$ avec $F(12, \infty; 0.05) = 1.753$ pour trois méthodes d'estimation: EMV, CPLX et TAYLOR.

Lorsqu'il y a corrélation interne, on observe une forte distorsion de l'erreur de première espèce avec la méthode EMV, même lorsque ζ^2 relativement faible ($\zeta^2 = 0.2$) pour des grappes de taille $m = 6$. Cette distorsion se voit par le biais relatif, qui varie en l'occurrence de 7 à 18. Ces valeurs représentent des erreurs de première espèce anormalement élevées, qui vont de 0.40 à 0.95. La méthode CPLX donne des biais relatifs acceptables même pour de petits échantillons. Quant à la méthode TAYLOR, elle produit des biais relatifs trop élevés dans le cas de petits échantillons. Toutefois, pour ce qui est des grands échantillons, elle équivaut à la CPLX. Celle-ci semble une fois de plus surpasser la méthode TAYLOR pour ce qui est des petits échantillons.

Tableau 3.5

Biais relatif de l'erreur de première espèce estimée pour le test F de $H_0: \hat{\beta} = \hat{\beta}^0$ avec un seuil nominal de 0.05 selon le plan d'échantillonnage II

Méthode					
n	ζ^2	ϕ	EMV	CPLX	TAYLOR
20	0.0	1	0.54	0.46	13.52
20	0.2	2	7.30	0.46	12.96
20	0.4	3	13.70	0.68	13.96
20	0.6	4	17.08	0.60	14.72
30	0.0	1	0.28	0.78	7.78
30	0.2	2	8.72	0.72	8.16
30	0.4	3	14.84	0.72	9.32
30	0.6	4	17.50	0.82	9.23
40	0.0	1	0.36	0.56	5.16
40	0.2	2	9.28	0.56	5.76
40	0.4	3	15.38	0.64	5.84
40	0.6	4	17.76	0.70	5.80
50	0.0	1	0.44	0.56	3.44
50	0.2	2	9.34	0.08	4.86
50	0.4	3	15.48	0.38	4.36
50	0.6	4	17.56	0.46	4.16
100	0.0	1	0.16	0.04	1.26
100	0.2	2	9.46	0.26	1.46
100	0.4	3	15.94	0.44	2.00
100	0.6	4	18.16	0.14	1.46
200	0.0	1	0.10	0.26	0.76
200	0.2	2	10.20	0.34	0.82
200	0.4	3	16.22	0.02	0.48
200	0.6	4	18.06	0.06	0.52

3.2 Plan d'échantillonnage II

Soit x_1, x_2, \dots, x_n une série de k -vecteurs aléatoires indépendants et identiquement distribués selon une loi normale avec une moyenne $\bar{\mu}$ et une matrice de covariances $\bar{\Sigma}_B$. Ces vecteurs représentent les moyennes de grappe pour les variables explicatives de la fonction logistique (2.1). Supposons que pour la grappe j , $j = 1, 2, \dots, n$, $x_{j0}^0, x_{j1}^0, \dots, x_{jm_j}^0$ sont des vecteurs aléatoires indépendants et identiquement distribués selon une loi normale avec une moyenne x_j et une matrice des covariances $\bar{\Sigma}_W$. Étant donné $x_{j0}^0, \ell = 0, 1, \dots, m_j$, le vecteur aléatoire $y_{j\ell}^0$ suit une distribution multinomiale avec paramètres $(\bar{\pi}_{j\ell}^0, 1)$, où les éléments de $\bar{\pi}_{j\ell}^0$ satisfont la fonction logistique (2.1) évaluée en fonction du vecteur de paramètres $\tilde{\theta}^0$ et à $x = x_{j0}^0$. De plus, supposons que les $x_{j\ell}^0$, sont indépendants, étant donné les $y_{j\ell}^0$.

Soit $U_{j1}, U_{j2}, \dots, U_{jm_j}$ variables aléatoires indépendantes et identiquement distribuées selon une loi uniforme $(0, 1)$, qui sont aussi conjointement indépendantes des $x_{j\ell}^0$ et des $y_{j\ell}^0$. Soit ζ un nombre fixe et connu, $0 \leq \zeta \leq 1$. Définissons maintenant $(x_{j\ell}, y_{j\ell}^*)$, $\ell = 1, 2, \dots$, m_j de la façon suivante:

$$(3.2.1) \quad (x_{j\ell}, y_{j\ell}^*) \equiv (x_{j0}^0, y_{j0}^0) \text{ si } U_{j\ell} \leq \zeta$$

et

$$(3.2.2) \quad (x_{j\ell}, y_{j\ell}^*) \equiv (x_{j\ell}^0, y_{j\ell}^0) \text{ si } U_{j\ell} > \zeta.$$

Notons que, dans chaque grappe, les $x_{j\ell}'$ ont tous le même vecteur de moyennes conditionnelles x_j et que la matrice des covariances de $x_{j\ell}$ et de $x_{j\ell}'$ est $\bar{\Sigma}_W$ si $\ell = t$ et $\bar{\zeta}^2 \bar{\Sigma}_W$ dans le cas contraire. Notons également que la moyenne conditionnelle de chaque $y_{j\ell}^*$ est la fonction logistique (2.1) évaluée à $\tilde{\theta}^0$ et $x = x_{j\ell}$ et que les vecteurs $(x_{j\ell}, y_{j\ell}^*)$, $\ell = 1, 2, \dots, m_j$, présentent une corrélation interne $\bar{\zeta}^2$ et un effet du plan approximatif $\phi = [1 + \bar{\zeta}^2(m - 1)]$ lorsque m_j est constant.

Des données $(x_{j\ell}, y_{j\ell}^*)$, $j = 1, 2, \dots, n$, $\ell = 1, 2, \dots, m_j$, ont été produites suivant le second plan d'échantillonnage en grappes pour $k = 4$, $d = 3$ et les paramètres

$$(3.2.3) \quad \bar{\mu} = (1, -6, 4, 8)',$$

$$(3.2.4) \quad \bar{\Sigma}_B = \text{Diag}(0, 25, 25, 49),$$

$$(3.2.5) \quad \bar{\Sigma}_W = \text{Diag}(0, 25, 36, 36),$$

$$(3.2.6) \quad \bar{\theta}_0^1 = (0.30, -0.05, -0.06, 0.08),$$

$$(3.2.7) \quad \bar{\theta}_0^2 = (0.06, -0.08, -0.10, 0.07),$$

et

$$(3.2.8) \quad \bar{\theta}_0^3 = (0.70, -0.08, -0.10, 0.11),$$

Tableau 3.4
Biais relatif des 5^e et 95^e percentiles estimés pour les variables "r" calculées pour les coefficients estimés, selon le plan d'échantillonnage I

Méthode					
n	\bar{z}^2	$\phi^{0.5} - 1$	EMV		CPLX
			5 ^e percentile	95 ^e percentile	
20	0.00	0.00	0.02	0.00	0.10
20	0.05	0.41	0.40	0.38	0.04
20	0.10	0.73	0.68	0.65	0.07
20	0.15	1.00	0.84	0.79	0.07
30	0.00	0.00	0.00	0.02	0.10
30	0.05	0.41	0.43	0.38	0.01
30	0.10	0.73	0.73	0.70	0.02
30	0.15	1.00	0.97	0.91	0.01
40	0.00	0.00	0.01	0.01	0.07
40	0.05	0.41	0.38	0.41	0.03
40	0.10	0.73	0.70	0.72	0.03
40	0.15	1.00	0.96	0.93	0.01
50	0.00	0.00	0.01	0.01	0.05
50	0.05	0.41	0.43	0.40	0.00
50	0.10	0.73	0.71	0.70	0.01
50	0.15	1.00	0.97	0.96	0.02
100	0.00	0.00	0.00	0.02	0.01
100	0.05	0.41	0.42	0.42	0.02
100	0.10	0.73	0.71	0.74	0.01
100	0.15	1.00	1.03	0.99	0.04
200	0.00	0.00	0.01	0.01	0.00
200	0.05	0.41	0.42	0.43	0.01
200	0.10	0.73	0.71	0.72	0.01
200	0.15	1.00	1.00	1.00	0.02
400	0.00	0.00	0.01	0.01	0.01
400	0.05	0.41	0.39	0.40	0.01
400	0.10	0.73	0.76	0.77	0.03
400	0.15	1.00	1.02	0.89	0.02
800	0.00	0.00	0.00	0.01	0.00
800	0.05	0.41	0.43	0.44	0.01
800	0.10	0.73	0.76	0.70	0.02
800	0.15	1.00	1.07	1.04	0.04

Tableau 3.3

Caractéristiques de ϕ selon le plan d'échantillonnage I de l'étude de Monte Carlo

Méthode						
n	ζ^2	ϕ	CPLX		TAYLOR	
			Biais rel.	E.T.	Biais rel.	E.T.
20	0.00	1	0.28	0.23	0.23	0.22
20	0.05	2	0.01	0.63	0.35	0.48
20	0.10	3	0.07	0.93	0.40	0.70
20	0.15	4	0.15	1.15	0.46	0.85
30	0.00	1	0.33	0.22	0.17	0.20
30	0.05	2	0.14	0.62	0.25	0.47
30	0.10	3	0.08	0.88	0.30	0.66
30	0.15	4	0.04	1.18	0.33	0.90
40	0.00	1	0.26	0.18	0.14	0.18
40	0.05	2	0.14	0.53	0.19	0.42
40	0.10	3	0.10	0.83	0.22	0.67
40	0.15	4	0.07	1.13	0.25	0.91
50	0.00	1	0.18	0.18	0.11	0.17
50	0.05	2	0.09	0.48	0.16	0.41
50	0.10	3	0.07	0.75	0.18	0.64
50	0.15	4	0.04	0.97	0.21	0.83
100	0.00	1	0.07	0.13	0.06	0.13
100	0.05	2	0.04	0.34	0.08	0.32
100	0.10	3	0.01	0.54	0.10	0.51
100	0.15	4	0.01	0.69	0.11	0.65
200	0.00	1	0.03	0.10	0.03	0.09
200	0.05	2	0.02	0.25	0.04	0.24
200	0.10	3	0.01	0.38	0.05	0.36
200	0.15	4	0.01	0.49	0.05	0.48
400	0.00	1	0.01	0.07	0.01	0.07
400	0.05	2	0.01	0.19	0.02	0.19
400	0.10	3	0.00	0.27	0.02	0.27
400	0.15	4	0.00	0.37	0.02	0.37
800	0.00	1	0.01	0.05	0.01	0.05
800	0.05	2	0.00	0.13	0.01	0.13
800	0.10	3	0.00	0.19	0.01	0.18
800	0.15	4	0.00	0.24	0.01	0.24

Tableau 3.2
Caractéristiques du critère chi carré utilisé pour le test de $H_0: \hat{\theta} = \theta_0$
selon le plan d'échantillonnage I de l'étude de Monte Carlo

Méthode			CPLX			TAYLOR		
n	$\hat{\epsilon}^2$	ϕ	Moyenne	Variance	Moyenne	Variance	Moyenne	Variance

20	0.00	1	11.5	22.2	12.0	32.7	81.9	12x10 ³
20	0.05	2	23.9	134.3	16.5	81.2	116.6	8x10 ⁴
20	0.10	3	34.2	239.9	16.6	77.8	94.5	12x10 ³
20	0.15	4	43.8	403.2	17.3	89.3	140.3	19x10 ⁴
30	0.00	1	11.8	25.1	11.2	28.5	35.1	702.3
30	0.05	2	23.8	121.4	13.2	41.2	34.1	691.6
30	0.10	3	35.8	268.1	13.8	46.3	41.2	12x10 ²
30	0.15	4	46.7	450.1	14.1	51.1	44.5	16x10 ²
40	0.00	1	12.2	24.3	11.9	30.3	25.8	268.3
40	0.05	2	23.2	96.5	12.6	33.6	25.4	201.4
40	0.10	3	35.4	247.7	13.5	43.3	29.1	340.4
40	0.15	4	46.2	428.9	13.8	44.4	30.2	331.4
50	0.00	1	11.9	25.5	12.4	34.6	21.0	140.8
50	0.05	2	23.9	112.5	13.7	43.8	22.7	153.6
50	0.10	3	35.8	231.0	14.3	46.0	24.6	195.8
50	0.15	4	46.7	424.0	14.5	55.4	25.2	234.6
100	0.00	1	12.1	23.6	13.2	35.0	15.8	55.0
100	0.05	2	23.9	102.6	13.8	39.2	16.5	62.1
100	0.10	3	36.5	233.9	14.6	47.0	17.6	75.8
100	0.15	4	47.5	350.4	14.6	43.0	17.9	70.6
200	0.00	1	11.7	24.1	12.6	32.4	13.6	38.2
200	0.05	2	23.9	93.9	13.1	33.1	14.1	39.1
200	0.10	3	35.7	194.1	13.3	31.5	14.3	37.4
200	0.15	4	48.0	399.6	13.5	35.7	14.6	42.7
400	0.00	1	11.9	24.9	12.3	29.3	12.7	31.3
400	0.05	2	24.1	96.6	12.7	29.2	13.1	31.3
400	0.10	3	36.9	208.5	13.1	29.2	13.6	31.4
400	0.15	4	47.3	390.7	12.7	31.6	13.1	34.0
800	0.00	1	11.9	24.0	12.1	26.4	12.3	27.2
800	0.05	2	24.0	99.3	12.3	27.3	12.5	28.2
800	0.10	3	36.4	239.3	12.6	30.1	12.8	31.1
800	0.15	4	48.7	396.3	12.6	26.7	12.7	27.5

En ce qui concerne la méthode TAYLOR, on observe un biais relatif élevé pour les petits échantillons. En effet, le biais relatif varie de 17 à 7 pour des échantillons dont la taille varie de 20 à 50. Pour de grands échantillons, les méthodes CPLX et TAYLOR donnent, comme prévu, des résultats similaires. En règle générale, les biais relatifs sont moins élevés dans le cas de la CPLX.

Si l'on multiplie la statistique F utilisée pour tester $H_0: \tilde{\beta} = \tilde{\beta}^0$ par le nombre de paramètres testés, on obtient une statistique qui est distribuée comme une variable aléatoire chi-carré avec 12 degrés de liberté. Le tableau 3.2 donne les moyennes et la variances de Monte Carlo pour cette statistique.

Comme prévu, la méthode EMV donne des moyennes qui se situent autour de 12 et des variances qui se situent autour de 24 lorsque l'effet du plan ϕ est égal à 1. Dans les mêmes conditions, la méthode CPLX donne des moyennes qui se situent autour de 12 et des variances qui sont plus élevées que dans le cas de l'EMV mais qui diminuent à mesure que la taille de l'échantillon augmente. En revanche, lorsqu'il y a corrélation interne, les moyennes et les variances obtenues par la méthode EMV sont excessives tandis que celles obtenues par la méthode CPLX sont en conformité avec la théorie asymptotique et la correction introduite en (2.13-2.14). La méthode TAYLOR donne des variances extrêmement élevées pour de petits échantillons. Cela s'expliquerait pas le fait que dans certaines répétitions de la simulation, la matrice des covariances (2.10) était mal-conditionnée, ce qui aurait donné une expression quadratique démesurée pour (2.11). Le problème s'atténue lorsque la taille de l'échantillon augmente. Les méthodes CPLX et TAYLOR deviennent asymptotiquement équivalentes pour de grands échantillons.

Le tableau 3.3 donne les caractéristiques de l'estimateur de l'effet du plan (3.1.8) selon l'étude de Monte Carlo pour les méthodes CPLX et TAYLOR. Les biais sont plus faibles mais les erreurs types légèrement plus élevées dans le cas de la méthode CPLX. Les deux méthodes donnent d'assez bons résultats.

Pour chaque catégorie r , $r = 1, 2, 3$ et chaque covariable s , $s = 1, 2, 3, 4$, nous avons aussi calculé des statistiques $''r''$ pour chacun des coefficients estimés en utilisant la formule suivante:

$$''r'' = [\text{Var}(\hat{\beta}_{rs})]^{-0.5}(\hat{\beta}_{rs} - \beta_{rs}^0). \tag{3.1.15}$$

Nous avons groupé les douze statistiques $''r''$ calculées à l'aide de la méthode CPLX et avons calculé ensuite les percentiles simulés. Nous avons répété ce calcul pour les statistiques $''r''$ calculées à l'aide de la méthode EMV. En conséquence, les percentiles reposent dans chaque cas sur 12 000 valeurs $''r''$. Une fois les percentiles calculés, nous avons estimé les biais relatifs au moyen de la formule suivante:

$$(\text{percentile normal type})^{-1} | \text{percentile estimé} - \text{Percentile normal type} |. \tag{3.1.16}$$

Le tableau 3.4 donne le biais relatif des 5^e et 95^e percentiles estimés pour la statistique $''r''$ calculée selon les deux méthodes. En ce qui concerne la méthode EMV, les biais relatifs devraient être proches de $\phi^{0.5} - 1$ puisque la statistique $''r''$ calculée selon cette méthode comporte un facteur $\phi^{0.5}$. Cette similitude ressort clairement dans le tableau 3.4 lorsque l'on regarde les deux colonnes de chiffres figurant sous EMV. Quant à la méthode CPLX, les biais relatifs sont acceptables pour de petits échantillons. Comme prévu, ces biais deviennent négligeables lorsque la taille de l'échantillon augmente.

Tableau 3.1

Biais relatif de l'erreur de première espèce estimée pour le test F de l'hypothèse $H_0: \hat{\beta} = \tilde{\beta}_0$ avec un seuil nominal de 0.05 selon le plan d'échantillonnage I

n	ξ^2	ϕ	Méthode	
			EMV	CPLX TAYLOR
20	0.00	1	0.24	0.60
20	0.05	2	9.66	3.68
20	0.10	3	15.24	3.98
20	0.15	4	17.74	4.00
30	0.00	1	0.08	0.06
30	0.05	2	9.84	1.20
30	0.10	3	15.52	1.76
30	0.15	4	17.74	1.86
40	0.00	1	0.04	0.32
40	0.05	2	9.98	0.82
40	0.10	3	16.20	1.02
40	0.15	4	17.74	1.80
50	0.00	1	0.06	0.50
50	0.05	2	9.76	1.44
50	0.10	3	16.00	1.96
50	0.15	4	17.80	2.20
100	0.00	1	0.06	0.90
100	0.05	2	10.02	1.66
100	0.10	3	16.26	2.06
100	0.15	4	17.78	2.24
200	0.00	1	0.02	0.74
200	0.05	2	10.46	1.00
200	0.10	3	16.30	0.88
200	0.15	4	18.00	1.52
400	0.00	1	0.02	0.44
400	0.05	2	10.14	0.66
400	0.10	3	16.56	0.64
400	0.15	4	17.86	0.56
800	0.00	1	0.08	0.32
800	0.05	2	10.36	0.22
800	0.10	3	16.04	0.68
800	0.15	4	18.12	0.50
800	0.40			0.54
800	0.36			0.80
400	0.70			0.90
400	0.90			1.00
400	1.00			0.84
200	1.28			1.64
200	1.64			1.88
200	2.12			2.12
100	2.68			3.90
100	3.90			4.70
100	5.10			5.10
50	7.40			8.38
50	8.38			9.32
50	9.70			9.70
40	9.66			9.66
40	9.62			9.62
40	11.66			11.66
30	12.82			13.74
30	13.74			14.22
30	14.68			14.68

où ζ^2 désigne la corrélation intra-grappe. De plus, si les m_j sont constantes, c.-à-d. $m_j = m$, le facteur $\phi = [1 + \zeta^2(m - 1)]$ sert à désigner l'effet du plan défini par Kish (1965, p.258). Un estimateur de l'effet du plan ϕ est

$$\phi = (dk)^{-1} \left[\sum_{d,k} a^{(i,t)} / h^{(i,t)} \right] w^{-1}, \tag{3.1.8}$$

où $a^{(i,t)}$ et $h^{(i,t)}$ représentent respectivement l'élément (i,t) de A^n dans (2.13) et (2.14) et l'élément (i,t) de $[H_n(\hat{g}^{\text{PSEUDO}})]^{-1}$, et w est la moyenne des poids d'échantillonnage pour tout l'échantillon.

Suivant ce plan d'échantillonnage, des données (x_j, y_j^*) , $j = 1, 2, \dots, n$, $\ell = 1, 2, \dots, m$, ont été produites pour $k = 4$, $d = 3$, $m = 21$ et les paramètres

$$\tilde{\mu} = (1, -2, 1, 5)', \tag{3.1.9}$$

$$\tilde{\Sigma} = \text{Diag}(0, 25, 25, 25), \tag{3.1.10}$$

$$\tilde{\beta}_0^1 = (-0.3, -0.1, 0.1, 0.2), \tag{3.1.11}$$

$$\tilde{\beta}_0^2 = (0.2, -0.2, -0.2, 0.1), \tag{3.1.12}$$

$$\tilde{\beta}_0^3 = (-0.1, 0.3, -0.3, 0.1). \tag{3.1.13}$$

et

À partir des cinq équations ci-dessus, on a produit 1000 séries d'échantillons avec n grappes de taille m suivant les équations (3.1.1) et (3.1.2) pour diverses valeurs de n , de ζ^2 et de ϕ . L'erreur de première espèce enregistrée par suite de la comparaison des résultats du test F de $H_0: \hat{\beta} = \hat{\beta}_0^a$ à $F(12, \infty; 0.05) = 1.753$ a été estimée selon les trois méthodes étudiées: EMV, CPLX et TAYLOR. Le biais relatif permet de mesurer l'écart entre la valeur estimée et la valeur nominale (0.05) de l'erreur de première espèce; ce biais est défini comme suit:

$$(0.05)^{-1} | \text{Erreur de première espèce estimée} - 0.05 |. \tag{3.1.14}$$

Le tableau 3.1 donne le biais relatif des erreurs de première espèce estimées. En l'absence de corrélation interne ($\zeta^2 = 0$), la méthode EMV donne, comme prévu, un biais relatif peu élevé. En revanche, ce biais est légèrement plus élevé avec la méthode CPLX. C'est l'inconvénient que présente l'estimation de paramètres additionnels dans les équations (2.13) et (2.14). Lorsqu'il existe une corrélation interne positive, la méthode EMV donne une erreur de première espèce estimée qui s'écarte sensiblement de la valeur nominale. Cet écart augmente avec le degré de corrélation interne ζ^2 . Lorsque $\zeta^2 = 0.15$ ($\phi = 4$), le biais relatif de l'erreur de première espèce estimée se situe autour de 18, ce qui signifie une erreur de première espèce anormalement élevée (environ .95). En ce qui a trait à la méthode CPLX, le biais relatif diminue à mesure que la taille de l'échantillon se rapproche du seuil critique de correction (2.14), qui est 34 en l'occurrence, puis augmente modérément à mesure que la taille de l'échantillon se rapproche de $n = 100$, pour ensuite diminuer progressivement pour des échantillons de taille supérieure à 100. Cette tendance est observée dans toute la simulation. Elle représente l'effet de la correction (2.13-2.14) dans les petits échantillons.

Contrairement au premier plan d'échantillonnage, le second prévoit des vecteurs de covariables différents pour chaque élément d'une grappe. La moyenne conditionnelle (2.1) est satisfaite et divers degrés de corrélation interne sont observés. Nous allons étudier l'effet de la corrélation interne pour les deux plans selon trois méthodes d'estimation: EMV, qui ne tient aucunement compte de l'effet de grappe; TAYLOR, qui utilise la matrice des covariances pour grand échantillon (2.10); et CPLX, qui utilise la matrice des covariances redressée (2.13-2.14). Dans le cas de grands échantillons, les deux dernières méthodes sont asymptotiquement équivalentes. Pour de petits échantillons, CPLX est supérieure à la méthode TAYLOR.

3.1 Plan d'échantillonnage I

Supposons que x_1, x_2, \dots, x_n sont des k -vecteurs aléatoires indépendants et identiquement distribués selon une loi normale avec une moyenne $\bar{\mu}$ et une matrice de covariances $\bar{\Sigma}$. Pour chaque $j, j = 1, 2, \dots, n$, supposons que, étant donné x_j , les vecteurs aléatoires $y_{j0}^0, y_{j1}^0, \dots, y_{jm_j}^0$ sont indépendants et identiquement distribués selon une loi multinomiale avec paramètres $(\bar{\pi}_j^*, 1)$, où $\bar{\pi}_j^*$ satisfait la fonction logistique (2.1) évaluée en fonction du vecteur de paramètres réel $\hat{\theta}_0^0$ et à $x = x_j$. Soit $U_{j1}, U_{j2}, \dots, U_{jm_j}$ un ensemble de variables aléatoires indépendantes et identiquement distribuées selon une loi uniforme $(0,1)$. Pour une valeur connue et fixe de $\zeta, 0 \leq \zeta \leq 1$, posons

(3.1.1)
$$y_{j\ell}^* \equiv y_{j0}^0 \quad \text{si} \quad U_{j\ell} \leq \zeta$$

et

(3.1.2)
$$y_{j\ell}^* \equiv y_{j\ell}^0 \quad \text{si} \quad U_{j\ell} > \zeta,$$

$\ell = 1, 2, \dots, m_j.$

Il est possible de montrer que dans la grappe j ,

(3.1.3)
$$E(y_{j\ell}^*) = \bar{\pi}_j^*,$$

(3.1.4)
$$\text{COV}(y_{j\ell}^*, y_{j\ell}^*) = \Delta(\bar{\pi}_j^*) \quad \text{si} \quad \ell = t,$$

et

(3.1.5)
$$\text{COV}(y_{j\ell}^*, y_{jt}^*) = \zeta^2 \Delta(\bar{\pi}_j^*) \quad \text{si} \quad \ell \neq t.$$

Par conséquent, étant donné x_j , le vecteur aléatoire $t_j = \sum_{\ell=1}^{m_j} y_{j\ell}^*$ ne suit pas une distribution multinomiale. Au lieu de cela,

(3.1.6)
$$E(m_j^{-1} t_j) = \bar{\pi}_j^*$$

et

(3.1.7)
$$\text{Var}(m_j^{-1} t_j) = [1 + \zeta^2(m_j - 1)] m_j^{-1} \Delta(\bar{\pi}_j^*),$$

La matrice des sommes des carrés et des sommes des produits utilisée dans la construction de G_n repose sur n observations (n = nombre de grappes). Par analogie avec la statistique T^2 d'Hotelling, il est naturel d'effectuer une correction comme celle qui consiste à multiplier l'expression (2.11) par le ratio

(2.12)

$$\frac{n}{n - v} v(n - 1)$$

pour obtenir un critère F approximatif avec v et $n - v$ degrés de liberté. Cette correction présente toutefois un inconvénient; en effet, il peut arriver que v soit supérieur à n lorsque nous avons peu de grappes mais beaucoup d'éléments dans l'échantillon. La matrice des covariances construite suivant l'hypothèse que les observations élémentaires constituent un échantillon aléatoire simple est biaisée mais elle peut servir à effectuer une correction pour petit échantillon dans la matrice des covariances estimée. On peut imaginer la correction habituelle pour petit échantillon, fondée sur les degrés de liberté, comme une opération qui consiste à ajouter la quantité $(n - v)^{-1} v V$, à V un estimateur initial de la matrice des covariances V étant par ailleurs un estimateur de cette matrice. Normalement, V est aussi l'estimateur initial de la matrice des covariances. Dans le cas qui nous occupe, nous faisons la correction à l'aide de la matrice des covariances formée des éléments de l'estimateur initial. Le fait d'utiliser la matrice des covariances élémentaire a pour avantage de produire une somme qui est toujours définie positive. La correction est fonction du nombre de paramètres estimés, dk . En conséquence,

(1) si $n > 3dk - 2$

(2.13)

$$\hat{A}_n = \hat{A}_n + (n - dk)^{-1} (dk - 1) \gamma^* [H_n(\hat{\theta}^{PSEUDO})]^{-1},$$

(2) si $n \leq 3dk - 2$

(2.14)

$$\hat{A}_n = \hat{A}_n + 0.5 \gamma^* [H_n(\hat{\theta}^{PSEUDO})]^{-1},$$

où $\gamma^* = \max\{1, \text{tr} \{ [H_n(\hat{\theta}^{PSEUDO})]^{-1} \hat{G}_n / dk \} \}$. La limite supérieure de 0.5 pour la correction dans l'équation (2.14) est arbitraire. On peut alors obtenir un test F approximatif avec v et $n - v$ degrés de liberté en substituant \hat{A}_n à \hat{A}_n dans l'équation (2.11) puis en divisant l'expression quadratique ainsi obtenue par v . En pratique, le nombre approximatif de degrés de liberté peut être v et l'infini.

3. ÉTUDE DE MONTE CARLO

Dans cette section, nous faisons une étude de Monte Carlo dans le but d'analyser les caractéristiques de tests F (2.11) qui font intervenir des paramètres de modèle. Les données sont produites suivant deux plans d'échantillonnage différents qui correspondent à un échantillonnage en grappes à un seul degré, où les unités primaires ont toutes le même poids d'échantillonnage et sont tirées d'une population infinie. Selon le premier plan d'échantillonnage, tous les éléments d'une grappe ont le même vecteur explicatif x et, partant, la même moyenne conditionnelle (2.1). C'est ce qu'on observe habituellement lorsque la régression logistique devient pondérée au sens où plusieurs réponses y ont le même vecteur de covariables x . Une corrélation interne s'établit à des degrés divers entre les y d'une même grappe.

Notons qu'un estimateur convergent de $H_n(\tilde{\beta}^0)$ est $H_n(\tilde{\beta}^{PSEUDO})$ et qu'un estimateur non paramétrique de G_n est

$$G_n^* = (n - 1)^{-1} \sum_{j=1}^n (d_j - \bar{d}), \tag{2.8}$$

où

$$d_j = \sum_{\ell=1}^{\ell} w_j(y_{j\ell} - \tilde{\pi}_{j\ell}) \otimes x'_{j\ell},$$

et $\bar{d} = n^{-1} \sum_{j=1}^n d_j$. Si, dans chaque grappe, les $y'_{j\ell}$ sont indépendants et identiquement distribués selon un vecteur aléatoire multinomial avec paramètres $(\tilde{\pi}'_j, 1)$, on peut montrer facilement que l'espérance de G_n^* est précisément $H_n(\tilde{\beta}^0)$. Dans la pratique, on remplace les $\tilde{\pi}_{j\ell}$ dans l'équation (2.8) par $\tilde{\pi}_{j\ell}'$, où $\tilde{\pi}_{j\ell}'$ est défini comme dans l'équation (2.1), $\tilde{\beta}^{PSEUDO}$ étant substitué à $\tilde{\beta}^0$, et on effectue une petite correction pour obtenir l'estimateur

$$\hat{G}_n = (n^* - k)^{-1} (n - 1)^{-1} \sum_{j=1}^n (d_j - \hat{\bar{d}}) (\hat{d}_j - \hat{\bar{d}})', \tag{2.9}$$

où

$$\hat{d}_j = \sum_{\ell=1}^{\ell} w_j(y_{j\ell} - \tilde{\pi}_{j\ell}) \otimes x'_{j\ell},$$

$$\hat{\bar{d}} = n^{-1} \sum_{j=1}^n \hat{d}_j \text{ and } n^* = \sum_{j=1}^n m_j.$$

Le facteur

$$(n^* - k)^{-1} (n - 1)^{-1} n$$

se ramène à $(n - k)^{-1} n$ si chaque grappe renferme exactement un élément. Le facteur $(n - k)^{-1} n$ représente un nombre de degrés de liberté et est le facteur de redressement appliqué au carré moyen des résidus calculé par la méthode des moindres carrés ordinaires où k paramètres sont estimés. La quantité exprimée par (2.9) est bien définie pour deux grappes ou plus et le facteur $(n^* - k)^{-1} (n - 1)$ devrait contribuer à réduire le biais dû aux petits échantillons, qui découle de l'utilisation de la fonction estimée dans le calcul des écarts. Par conséquent, un estimateur convergent de la matrice des covariances asymptotique de $\tilde{\beta}^{PSEUDO}$ selon un échantillonnage en grappes est

$$A_n = [H_n(\tilde{\beta}^{PSEUDO})]^{-1} G_n [H_n(\tilde{\beta}^{PSEUDO})]^{-1} \tag{2.10}$$

Cette formule peut servir à tester l'hypothèse du genre $H_0: C \tilde{\beta}^0 = \tilde{\delta}^*$. Selon l'hypothèse nulle (Moore, 1977),

$$(C \tilde{\beta}^{PSEUDO} - \tilde{\delta}^*)' [C A_n C']^{-1} (C \tilde{\beta}^{PSEUDO} - \tilde{\delta}^*) \tag{2.11}$$

converge en loi vers une distribution chi carré avec v degrés de liberté ($v = \text{rang}(C A_n C')$). Dans l'équation ci-dessus, $[C A_n C']^{-1}$ est n'importe quelle inverse généralisée de $C A_n C'$.

ou,

$$\begin{aligned}
 H^n(\tilde{\theta}^0) &= \sum_n \sum_{f=1}^F w_f \Delta(\tilde{\pi}_{f\ell}) \otimes x_{f\ell}^T x_{f\ell}, \\
 U^n(\tilde{\theta}^0) &= \sum_n \sum_{f=1}^F w_f (y_{f\ell} - \tilde{\pi}_{f\ell}) \otimes x_{f\ell}^T, \\
 G^n &= \sum_n \sum_{f=1}^F w_f^2 \text{Var}(y_{f\ell}) \otimes x_{f\ell}^T x_{f\ell},
 \end{aligned}$$

$y_{j\ell}$ et $\tilde{\pi}_{j\ell}$ étant les vecteurs $y_{j\ell}^*$ et $\tilde{\pi}_{j\ell}^*$ amputés de leur dernier élément et N^{dk} désignant une distribution normale à dk dimensions.

Nelder et Wedderburn (1972) ont montré que, dans l'hypothèse d'une distribution binomiale, il est possible de résoudre la pseudo-fonction de vraisemblance logarithmique (2.2) à l'aide d'une méthode des moindres carrés pondérés itérative. Haberman (1974, p. 48) montre que, dans des conditions de régularité, un algorithme de Newton-Raphson modifié converge vers l'estimateur du maximum de vraisemblance dans le cas d'une distribution multinomiale. Sa démonstration ne repose aucunement sur l'existence d'un estimateur convergent de $\tilde{\theta}^0$ qui permet l'initialisation de l'algorithme d'itération à $\tilde{\theta} = 0$. Jennrich et Moore (1975) ont démontré que l'algorithme de Gauss-Newton, couramment utilisé pour déterminer l'estimateur du maximum de vraisemblance de $\tilde{\theta}^0$ devient l'algorithme de Newton-Raphson lorsque l'hypothèse de la distribution multinomiale est valide. Comme il y a équivalence entre ces algorithmes et qu'un algorithme de Newton-Raphson modifié converge toujours, nous avons adopté la version modifiée de l'algorithme de Gauss-Newton décrite par Gallant (1987, p. 318).

Selon la méthode CPLX, on détermine tout d'abord $\tilde{\theta}^{\text{PSEUDO}}$ au moyen d'un processus itératif selon lequel la valeur estimée de $\tilde{\theta}^0$ à la q -ième itération est

$$\tilde{\theta}^{[q, i(q)]} = \tilde{\theta}^{[q-1, i(q-1)]}$$

$$+ (0.5)^{i(q)} [H^n(\tilde{\theta}^{[q-1, i(q-1)]})]^{-1} U^n(\tilde{\theta}^{[q-1, i(q-1)]}) \quad (2.5)$$

où $i(q)$ est un entier non négatif de telle sorte que

$$L^n(\tilde{\theta}^{[q, i(q)]}) > L^n(\tilde{\theta}^{[q-1, i(q-1)]}). \quad (2.6)$$

La modification de l'algorithme d'itération représentée par $i(q)$ assure la convergence du processus. On déclenche l'itération en posant $\tilde{\theta}^{(0)} = 0$. L'algorithme est réputé pour avoir convergé lorsque la condition

$$\frac{L^n(\tilde{\theta}^{[q, i(q)]}) - L^n(\tilde{\theta}^{[q-1, i(q-1)]})}{|L^n(\tilde{\theta}^{[q, i(q)]})| + 10^{-5}} < \epsilon \quad (2.7)$$

est satisfaite, où ϵ peut être 10^{-8} .

celui-là, sera égal à un. Soit $x_{j\ell}$ un vecteur ligne à k dimensions, constitué de variables explicatives et rattaché à l'unité ℓ tirée de la grappe j . Alors, pour chaque $j = 1, 2, \dots, n$, et chaque $\ell = 1, 2, \dots, m_j$, l'espérance de l'élément r de $y_{j\ell}^*$ est déterminée par une relation logistique définie

$$\pi_{j\ell r} = E\{y_{j\ell r}\} = [1 + \sum_{s=1}^S \exp(x_{j\ell} \tilde{\beta}_s^0)]^{-1} \exp(x_{j\ell} \tilde{\beta}_r^0) \quad r = 1, 2, \dots, d$$
$$= 1 - \sum_{p \neq r} \pi_{j\ell s}, \quad r = d + 1. \tag{2.1}$$

Comme la fonction d'espérance est non linéaire par rapport au vecteur de paramètres $\tilde{\beta}^0 = (\tilde{\beta}_1^0, \tilde{\beta}_2^0, \dots, \tilde{\beta}_d^0, \tilde{\beta}_{d+1}^0, \dots, \tilde{\beta}_S^0)$, il faut recourir à des méthodes d'estimation non linéaires. Définissons la pseudo-fonction de vraisemblance logarithmique $L_n(\tilde{g})$ comme suit:

$$L_n(\tilde{g}) = \sum_n \sum_{m_j} w_j (\log \pi_{j\ell}^*)' y_{j\ell}^*, \tag{2.2}$$

où $\pi_{j\ell}^* = (\pi_{j\ell 1}, \dots, \pi_{j\ell, d+1})'$ et w_j est le poids d'échantillonnage rattaché à l'unité d'échantillonnage $j\ell$. On peut considérer l'expression ci-dessus comme une fonction de vraisemblance logarithmique pondérée, où les poids sont les poids d'échantillonnage et les $y_{j\ell}^*$ sont distribués comme des variables aléatoires multinomiales. Si les poids d'échantillonnage sont tous égaux à un, (2.2) devient la fonction de vraisemblance logarithmique suivant l'hypothèse que les $y_{j\ell}^*$ sont indépendants et identiquement distribués selon une loi multinomiale. Définissons \hat{g}_{PSEUDO} comme l'estimateur de \tilde{g}^0 qui maximise (2.2). Cet estimateur représente une solution au système d'équations

$$\sum_n \sum_{m_j} w_j G(\tilde{g}, x_{j\ell}) [\text{Diag}(\pi_{j\ell}^*)]^{-1} (y_{j\ell}^* - \pi_{j\ell}^*) = 0, \tag{2.3}$$

où

$$G(\tilde{g}, x_{j\ell}) = [I^{d \times d}, 0^{d \times 1}] \Delta(\pi_{j\ell}^*),$$
$$\Delta(\pi_{j\ell}^*) = \text{Diag}(\pi_{j\ell}^*) - \pi_{j\ell}^* (\pi_{j\ell}^*)',$$

et \otimes désigne le produit tensoriel. On peut démontrer la normalité asymptotique de \hat{g}_{PSEUDO} en définissant implicitement les paramètres d'intérêt, comme dans l'équation (2.2), puis en donnant une portée plus générale aux résultats rapportés dans Binder (1983). On peut aussi procéder en se servant de l'hypothèse de pseudo-vraisemblance et de la Proposition 1 énoncée dans Dale (1986). Binder et Dale posent tous deux les conditions de régularité nécessaires. Lorsque n augmente,

$$\sqrt{n}(\hat{g}_{\text{PSEUDO}} - \tilde{g}^0) = \sqrt{n}[H_n(\tilde{g}^0)]^{-1} U_n(\tilde{g}^0)$$
$$\xrightarrow{N^{dk}} (0, \lim_{n \rightarrow \infty} [H_n(\tilde{g}^0)]^{-1} G_n[H_n(\tilde{g}^0)]^{-1}) \tag{2.4}$$

de régression logistique et des modèles discontinus de hasards proportionnels. Albert et Lesaffre (1986) ont analysé la méthode de discrimination logistique, par laquelle on attribue des observations multidimensionnelles à une population parmi plusieurs. Ils concentrent leur attention sur les groupes qualitativement différents.

Bull et Pederson (1987) et Morel (1987) ont approfondi le cas où la réponse consiste en une variable polychotomique. En se servant du développement de Taylor, ils montrent que la variance des estimations beta (pour grands échantillons) s'écrit

$$H^{-1}GH^{-1}$$

où H^{-1} est la matrice des covariances qui découle indûment des hypothèses de l'indépendance et de la distribution multinomiale appliquées aux éléments du vecteur de réponse, et G est une matrice dont l'estimation repose sur le plan de sondage complexe.

Plus récemment, Roberts, Rao et Kumar (1987) ont montré comment effectuer des corrections qui tiennent compte du plan de sondage lorsqu'on calcule le critère chi carré et le critère du rapport des vraisemblances dans l'analyse de régression logistique avec variable de réponse binaire. Ces corrections reposent sur certains effets du plan généralisés. On peut appliquer les résultats de Roberts, Rao et Kumar (1987) au cas de la population qui est divisée en I domaines d'étude, pour chacun desquels on doit prélever un grand échantillon et estimer une proportion $\pi_i, i = 1, 2, \dots, I$. On suppose que

$$\pi_i = [1 + \exp(x_i' \tilde{\theta}^0)]^{-1} \exp(x_i' \tilde{\theta}^0), i = 1, 2, \dots, I,$$

où x_i est un k -vecteur de constantes connues tirées du domaine i et $\tilde{\theta}^0$ est un k -vecteur de paramètres inconnus. Cette méthode est surtout utile lorsqu'on dispose du tableau récapitulatif des totaux et des facteurs de redressement de la variance au lieu de la série de données complète. Cet article a pour but d'exposer une méthode qui permet de calculer des estimateurs convergents du vecteur de paramètres d'un modèle logistique généralisé et de la matrice des covariances asymptotique correspondante dans le cas d'un plan de sondage complexe. La matrice des covariances estimée est toujours définie positive et asymptotiquement équivalente à celle obtenue par le développement de Taylor. Il sera aussi question dans cet article d'une correction permettant de réduire le biais dû aux petits échantillons dans la matrice des covariances estimée. Nous allons montrer par une étude de Monte Carlo que cette correction ramène à un niveau plus acceptable l'erreur de première espèce anormalement élevée qui se produit lorsqu'on ne tient pas compte du plan de sondage complexe, et ce plus rapidement que ne le fait le développement de Taylor. En ce sens, nous pouvons dire que cette correction donne, pour de petits échantillons, des résultats supérieurs à ceux que l'on obtient par la méthode delta habituelle. Nous allons désigner la nouvelle méthode par le terme CPLX. La méthode du maximum de vraisemblance et la méthode du développement de Taylor seront désignées respectivement par EMV et TAYLOR. La méthode CPLX a été incorporée à PC CARP, qui est un programme d'ordinateur personnel servant à l'estimation de la variance pour de grandes enquêtes (voir Schnell et coll. 1988).

2. RÉGRESSION LOGISTIQUE AVEC ÉCHANTILLONNAGE EN GRAPPES

Considérons d'abord un échantillonnage en grappes à un seul degré, où n grappes ou unités primaires d'échantillonnage sont prélevées avec remise (et avec probabilité connue) dans une population finie ou sans remise dans une très grande population. Soit m_j la taille de la grappe $j, j = 1, 2, \dots, n$, et y_{ji}^* , $i = 1, 2, \dots, m_j$, des vecteurs de classification à $(d + 1)$ dimensions. Le vecteur y_j^* est composé entièrement de zéros sauf lorsque l'unité i tirée de la grappe j appartient à la catégorie r ; dans ces circonstances, l'élément r du vecteur, et uniquement

Régression logistique selon des plans de sondage complexes

JORGE G. MOREL¹

RÉSUMÉ

L'auteur expose des méthodes qui permettent de calculer des estimateurs convergents des paramètres d'une fonction logistique générale et de la matrice des covariances asymptotique correspondante selon des plans de sondage complexes. Il corrige l'estimateur de Taylor de la matrice des covariances de manière à obtenir une matrice définie positive et à réduire du même coup le biais dû aux petits échantillons. L'auteur s'intéresse tout d'abord au cas de l'échantillonnage en grappes et passe ensuite à des plans de sondage plus complexes. Il réalise une étude de Monte Carlo afin d'analyser les caractéristiques de tests F construits à partir de diverses matrices de covariances pour de petits échantillons. Enfin, il compare la méthode du maximum de vraisemblance, qui ne tient aucunement compte du plan de sondage, avec la méthode du développement de Taylor et une version modifiée de celle-ci.

MOTS CLÉS: Pseudo-vraisemblance; méthode CPLX; échantillonnage en grappes; matrice des covariances redressée.

1. INTRODUCTION

Depuis quelques années, les statisticiens s'intéressent de près aux problèmes qui surgissent lorsque des données obtenues à l'aide de plans de sondage complexes font l'objet de tests chi carré fondés sur la distribution multinomiale. Ils ont montré que les effets de la stratification et de la formation de grappes sur ce genre de tests pouvaient amener une distorsion des niveaux de signification nominaux. Holt, Scott et Ewings (1980) ont proposé une version modifiée de tests chi carré de validité de l'ajustement, d'homogénéité et d'indépendance dans les tableaux à deux dimensions. Rao et Scott (1981) ont présenté des tests semblables pour des enquêtes à plan de sondage complexe. Dans tous ces cas, les estimations de variance (ou les effets du plan) pour chaque case suffisent pour déterminer le facteur de redressement. Bedrick (1983) a calculé un facteur de redressement pour tester l'ajustement de modèles log-linéaires hiérarchiques à l'aide d'estimations de paramètres en forme analytique. Par la suite, Rao et Scott (1984) ont présenté des façons plus étoffées d'utiliser les effets du plan pour définir des tests chi carré pour les enquêtes à plan de sondage complexe. Ils ont étendu les résultats qu'ils avaient obtenus antérieurement aux tableaux à plusieurs dimensions. Fay (1985) a présenté les corrections qu'il avait apportées au critère chi carré de Pearson et au critère du rapport des vraisemblances à l'aide d'une méthode "jackknife".

Le modèle logistique conditionnel (Cox 1970) est de plus en plus utilisé pour les plans de sondage complexes. Binder (1983) a démontré, dans des conditions favorables, la normalité asymptotique de la distribution d'échantillonnage fondée sur un plan pour une famille d'estimateurs de paramètres que l'on ne peut définir explicitement comme une fonction d'autres statistiques tirées de l'échantillon. Il applique les résultats de son analyse à des modèles logistiques binaires. On trouve aussi des applications de ce modèle à l'Enquête Santé Canada dans Binder et coll. (1984).

Chambliss et Boyle (1985) ont élaboré une théorie générale de la distribution asymptotique pour des échantillons aléatoires stratifiés avec un nombre fixe de strates et des tailles d'échantillon de strate croissantes. Ils ont illustré leurs résultats théoriques par des modèles

¹ Jorge G. Morel est professeur adjoint au Département d'épidémiologie et de biostatistique de l'University of South Florida, Tampa, Floride 33612.

- HIDIROGLOU, M.A., et RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: Part I - simple goodness-of-fit, homogeneity and independence in a two-way table with applications to the Canada Health Survey (1978-1979). *Journal of Official Statistics*, 3, 117-132.
- HOCHBERG, Y., et TAMANE, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
- HOLT, D., SCOTT, A.J., et EWING, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Sér. A*, 143, 303-320.
- JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *The American Statistician*, 41, 335-337.
- MILLER, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Edition. New York: Springer-Verlag.
- QUESENBERRY, C.P., et HURST, D.C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6, 191-195.
- RAO, J.N.K., et SCOTT, A.J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 58-66.
- RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 261-230.
- RAO, J.N.K., et THOMAS, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.
- THOMAS, D.R. (1989). An investigation of simultaneous confidence interval procedures for proportions under cluster sampling. Document de travail WPS 89-02, School of Business, Carleton University.
- THOMAS, D.R., et RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

lorsque r diminue et le conservatisme inhérent aux intervalles de type Scheffé) s'appliqueront de façon générale, on devrait pouvoir étendre à de nombreux plans de sondage les tendances qualitatives concernant les diverses statistiques analysées, même lorsque le nombre de grappes est peu élevé. Cette étude avait pour but fondamental de définir des méthodes de construction d'ICS où le taux d'erreur était peu influencé par les variations des paramètres à l'étude, soit le nombre de catégories, le nombre de grappes, le degré d'échantillonnage en grappes et l'asymétrie du vecteur des probabilités par catégorie. Comme les combinaisons de paramètres étudiées représentent un large éventail des cas que l'on peut rencontrer dans la pratique, il est raisonnable de croire que la robustesse observée dans le cas des intervalles de Bonferroni soumis aux transformations log et logit pourrait aussi s'appliquer aux variations de plan de sondage pour un nombre de grappes (ou de degrés de liberté) modéré. Il faut manifestement poursuivre les recherches sur cette question.

Sous réserve des commentaires ci-dessus, les intervalles de Bonferroni fondés sur le critère t et soumis à la transformation logit sont recommandés pour évaluer jusqu'à $k = 12$ proportions d'ordre de grandeur varié dans des conditions réalistes d'échantillonnage en grappes. Si l'on juge que le conservatisme est un atout, alors on peut utiliser sans problème les intervalles de Quesenberry-Hurst modifiés du premier degré. Dans les deux cas, il suffit de connaître les variances (ou les effets du plan) des proportions de cases estimées.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à J.N.K. Rao pour les nombreux échanges fructueux qu'il a eus avec lui et pour les commentaires que M. Rao a bien voulu faire sur une version préliminaire de l'article. L'auteur souhaite aussi remercier Steve Brockwell pour ses suggestions utiles et l'excellent appui qu'il a offert pour la programmation, ainsi que Paul Bertelman, qui a contribué à la programmation dans la dernière phase du projet et enfin les deux réviseurs et un rédacteur associé pour leurs suggestions qui ont permis d'améliorer sensiblement la présentation de l'article et la discussion des résultats. Cette étude a été rendue possible grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

ANDREWS, R.W., et BIRDSALL, W.C. (1988). Simultaneous confidence intervals: a comparison under complex sampling. Article présenté à la 1988 American Statistical Association Annual Meeting, Chicago.

BAILEY, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformation of the cell frequencies. *Technometrics*, 22, 583-589.

BLACK, D., et MYLES, J. (1986). Dependent industrialization and the Canadian class structure: a comparative analysis of Canada, the United States, and Sweden. *Canadian Review of Sociology and Anthropology*, 23, 157-181.

BRIER, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67, 591-596.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

FITZPATRICK, S., et SCOTT, A.J. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82, 875-878.

GOLD, R.Z. (1963). Tests auxiliary to χ^2 tests in a Markov chain. *Annals of Mathematical Statistics*, 34, 56-74.

GOODMAN, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.

8. RÉSUMÉ, CONCLUSIONS ET RECOMMANDATIONS

Dans cet article, nous avons tenté de trouver des méthodes de construction d'intervalles de confiance simultanés qui tiennent compte directement du plan de sondage et qui permettent de maintenir à un niveau acceptable les taux d'erreur et le degré d'asymétrie de ces taux dans diverses conditions d'échantillonnage en grappes. À cet égard, il faut rejeter les intervalles de Scheffé fondés sur les variances par case estimées: la version chi carrée de ces intervalles (équation 3) doit être rejetée à cause d'un taux d'erreur trop élevé et la version F doit l'être à cause d'une trop grande asymétrie. Les intervalles de Quesenberry-Hurst modifiés sont assez pratiques, quoiqu'ils obtiennent par la correction du premier degré de Rao-Scott représentent des ICS acceptables. En ce qui concerne les intervalles de Bonferroni, le fait d'utiliser les valeurs critiques de la distribution t plutôt que celles de la distribution centrée réduite procure des avantages substantiels. Malgré cela, les intervalles basés sur π et sa racine carrée ne présentent pas un taux d'erreur global acceptable, surtout pour de faibles valeurs de r , lorsque la distribution de π s'éloigne de la distribution normale. Par ailleurs, les intervalles de Bonferroni fondés sur le critère t et soumis aux transformations log et logit présentent un taux d'erreur global acceptable et un degré d'asymétrie des taux d'erreur très satisfaisant et sont, en définitive, nettement supérieurs à tous les autres types d'intervalles. Les intervalles de Bonferroni fondés sur le critère t (versions log et logit) semblent aussi présenter des taux d'erreur acceptables et un degré d'asymétrie très modéré lorsque les probabilités par case sont inégales, le rapport entre la probabilité la plus élevée et la plus faible pouvant aller jusqu'à seize, selon les cas analysés. Du point de vue du taux d'erreur global, il y a peu de différence entre les intervalles log et logit, sauf peut-être que les taux d'erreur sont généralement un peu moins élevés dans le second cas. Toutefois, pour ce qui est de la symétrie, les intervalles logit sont supérieurs aux intervalles log. Les valeurs estimées de longueurs d'intervalles de confiance (résultats détaillés non reproduits ici) tendent aussi à confirmer la supériorité des intervalles logit, malgré des taux d'erreur légèrement inférieurs. Par exemple, pour le cas d'équivalence de confiance pour π_1 (intervalle à 95 %) était .1915 \pm .0014 dans le cas de l'intervalle obtenu par transformation logarithmique, et .1850 \pm .0014 dans le cas de l'intervalle obtenu par transformation logit. Pour le cas des probabilités inégales, avec $\alpha = 5\%$, $k = 8$, $\lambda = 2$, $a = 0.71$, $r = 50$, $\pi_1 = 0.65$ et $\pi_2 = 0.05$ (voir tableau 5), les intervalles de confiance à 95 % pour la longueur moyenne des intervalles log et logit étaient, pour π_1 , $2.865 \pm .0012$ et $2.776 \pm .0011$, respectivement, et pour π_2 , $0.806 \pm .0010$ et $0.789 \pm .0011$, respectivement. Avant de passer aux recommandations finales, il faut considérer les lacunes que peut présenter le plan de l'étude de Monte Carlo. L'utilisation d'un plan d'échantillonnage unique, en l'occurrence l'échantillonnage en grappes à deux degrés avec EAS au second degré, pourrait constituer une lacune étant donné que les expérimentateurs auront à traiter des données qui auront été recueillies au moyen de plans de sondage variés prévoyant une stratification et un échantillonnage à plusieurs degrés. Pour de grands échantillons, la théorie de la distribution pertinente dira qu'il suffit de connaître les premier et second moments en supposant qu'un théorème limite central approprié s'applique (voir par exemple Rao et Scott 1981). Cette étude donnera donc lieu à des recommandations valables pour les cas où le nombre de grappes est élevé ou, de façon plus générale, pour les cas où le nombre de degrés de liberté pour l'estimation de la variance est élevé (Rao et Thomas 1988), pourvu que la matrice des covariances V/n et, par conséquent, les effets du plan généralisés puissent être modélisés convenablement. Comme le modèle de Dirichlet utilisé dans cette étude donne des effets du plan généralisés dont la moyenne et le coefficient de variation sont très fréquents en pratique, nous pouvons faire avec confiance des recommandations qui reposent sur un nombre élevé de grappes ou de degrés de liberté (cinquante ou plus). Lorsque le nombre de grappes est faible ou modéré, les résultats quantitatifs peuvent varier d'un plan à l'autre. Comme les mécanismes qui sont à la base des résultats présentés dans cette étude (notamment l'accentuation de la non-normalité de π

Tableau 5
Effet des probabilités de case inégales sur le taux d'erreur global (ER_T) et le degré d'asymétrie (PER_U) des intervalles Q-H modifiés et des intervalles de Bonferroni transformés;
 $r = 50, \lambda = 2, a = 5\%, m = 10$

Intervalles					
Q-H modifiés (premier degré)			Bonferroni fondés sur le critère t (logit)		
k	$\pi(k, q, \phi)$	ER_T	PER_U	ER_T	PER_U
5	$\pi(5, 1, 0, 8)$	3.2	7.3	5.6	75.9
5	$\pi(5, 2, 0, 425)$	1.4	82.1	4.8	57.2
5	$\pi(5, 3, 0, 3)$	1.5	76.7	4.2	51.2
5	equi-prob.	2.0	45.0	4.5	61.1
8	$\pi(8, 1, 0, 65)$	2.7	63.0	6.3	68.3
8	$\pi(8, 2, 0, 35)$	0.6	83.3	4.9	58.2
8	$\pi(8, 3, 0, 25)$	0.7	100	5.2	68.2
8	equi-prob.	2.5	66.5	6.0	64.0
49.0					

En ce qui concerne l'intervalle Q-H modifié, on constate que l'absence d'équiprobabilité influe réellement sur le taux d'erreur global, surtout lorsque $k = 8$. Dans le cas où $\pi_1 = 0.65$, le taux d'erreur global de l'intervalle Q-H modifié est voisin du taux d'erreur observé en situation d'équiprobabilité. Pour ce qui est des deux autres cas analysés ($\pi_1 = \pi_2 = .35$, et $\pi_1 = \pi_2 = 0.25$), le taux d'erreur global est beaucoup moins élevé; de fait, il se rapproche plus du taux observé pour le cas où les effets du plan sont constants (voir Thomas 1989). Si les taux d'erreur observés diffèrent les uns des autres, c'est que la structure des effets du plan par case varie d'un cas à l'autre même si la structure des effets du plan généralisés (λ) ne changent pas ($\lambda_1 = 2 + 2\sqrt{3}/3, \lambda_j = 2 - \sqrt{3}/3, j = 2, \dots, 7$ pour $\lambda = 2, a = \sqrt{2}/2 = .707$). L'utilisation d'un facteur de redressement uniforme (λ) aura donc pour conséquence de sous-estimer largement la variance de la première probabilité de case estimée, ce qui entraînera une augmentation du taux d'erreur de l'intervalle Q-H modifié. Les taux d'erreur observés sont tous au-dessus du taux nominal ($\alpha = 5\%$) à cause du conservatisme inhérent aux intervalles Q-H modifiés lorsque les effets du plan sont constants (voir section 5.3). Lorsque $\pi_1 = \pi_2 = 0.35$, les effets du plan correspondants sont $d_1 = d_2 = 2.36, d_i = 1.97, i = 3, \dots, 8$. Ces valeurs étant beaucoup plus proches de celle des effets du plan constants ($d_i = 2.0, i = 1, \dots, 8$) il n'est pas surprenant de constater un fort conservatisme des intervalles dans ce cas. On remarque aussi d'après le tableau 5 qu'une valeur de ER_T relativement plus faible va de pair avec un degré d'asymétrie relativement plus élevé.

Malgré la variation des effets du plan par case que supposent les différents vecteurs de probabilité du tableau 5, nous pouvons constater que les intervalles de Bonferroni transformés affichent des résultats très stables. Qu'il s'agisse des intervalles log ou logit, le taux d'erreur global (pour 50 grappes) est proche du niveau nominal ($\alpha = 5\%$) et le degré d'asymétrie est très modéré. Dans le cas de probabilités inégales, le taux d'erreur global diminue avec r (pour $r = 50$ à $r = 15$) lorsque $k = 8$ (résultats non reproduits ici). Toutefois, les variations de ER_T sont peu prononcées; lorsque $r = 15$ grappes, le taux d'erreur le plus bas pour les cas analysés est d'environ 2 %.

6. SYMÉTRIE DES TAUX D'ERREUR POUR LES

MÉTHODES ACCEPTABLES

Pour présenter les résultats relatifs à la symétrie des taux d'erreur, nous allons décomposer le taux d'erreur global ER_T en ces deux éléments ER_U et ER_L , comme nous l'avons décrit dans la section 4. La mesure utilisée dans les tableaux est $(ER_U/ER_T) \times 100\%$, c.-à-d. le taux d'erreur supérieur exprimé en pourcentage du taux d'erreur global. Désignons cette mesure par PER_U . Pour un ICS symétrique, on aura une valeur empirique de PER_U située autour de 50%; une valeur PER_U supérieure (inférieure) à 50% indiquera une probabilité accrue d'exclusion à cause d'intervalles situés à droite (à gauche) des valeurs π_i respectives. Pour des pourcentages de symétrie de 50 à 80%, l'intervalles de confiance à 95% pour la valeur réelle de PER_U est approximativement $(PER_U \neq 14\%)$ pour un taux d'erreur global de 5% et $(PER_U \neq 10\%)$ pour un taux d'erreur global de 10%.

6.1 Intervalles de Schéffé et de Quesenberry-Hurst modifiés

Le tableau 4 donne la symétrie (en pourcentage) du taux d'erreur global pour les intervalles de Schéffé fondés sur le critère F et les intervalles de Quesenberry-Hurst (Q-H) du premier degré pour une série de valeurs de paramètres. Il est facile de voir que l'intervalles de Schéffé est fortement asymétrique, ce qui en fait un ICS peu intéressant. L'intervalles Q-H modifié du premier degré n'est que modérément asymétrique et est donc préféré au premier dans la pratique. Le caractère asymétrique des intervalles de Schéffé est lui aussi attribuable à la non-normalité des π_i non transformés. En particulier, le fait que de "faibles" valeurs de π_i produisent de "faibles" estimations de la variance v_{ii} et, partant, des intervalles moins étendus (voir le cas multinomial où $v_{ii} = \pi_i(1 - \pi_i)n$, $i = 1, \dots, k$) accroît la probabilité que des intervalles se trouvent à gauche des valeurs π_i respectives. Cette tendance à l'asymétrie s'accroît à mesure que diminue le taux d'erreur global, ce qui rend l'intervalles de Schéffé fondé sur le critère F peu intéressant à ce point de vue. Comme les intervalles de Schéffé ne diffèrent des intervalles de Bonferroni simples que par la valeur critique utilisée, ces derniers devaient aussi être asymétriques, quoique dans une moins large mesure étant donné des taux d'erreur appréciables. C'est ce que viennent confirmer des résultats de l'étude (par exemple, $PER_u = 4.9\%$ pour des intervalles de Bonferroni simples fondés sur le critère t lorsque $r = 50$, $k = 8$ et $\alpha = 0.71$.

6.2 Intervalles de Bonferroni transformés fondés sur le critère t

Le tableau 4 donne également la symétrie (en pourcentage) du taux d'erreur global pour les intervalles de Bonferroni fondés sur le critère t et soumis aux transformations log et logit. Les résultats permettent de croire que les intervalles logit sont plus symétriques que les intervalles log pour des valeurs de k allant de 5 à 8. Par conséquent, du point de vue de la symétrie du taux d'erreur, les intervalles logit devraient être préférés aux intervalles log dans la pratique.

7. PROBABILITÉS DE CASE INÉGALES

Le tableau 5 donne le taux d'erreur global et le degré de symétrie du taux d'erreur observé pour les versions log et logit des intervalles de Bonferroni fondés sur le critère t et les intervalles Q-H modifiés du premier degré lorsque les probabilités de case sont inégales. Les résultats portent sur six cas de probabilités inégales, trois pour $k = 5$, $\lambda = 2$, $\alpha = 0.5$, notamment $\pi(5, 3, .3)$, $\pi(5, 2, .425)$ et $\pi(5, 1, .8)$, (voir section 4.1), et trois pour $k = 8$, $\lambda = 2$, $\alpha = 0.71$, notamment $\pi(8, 3, .25)$, $\pi(8, 2, .35)$ et $\pi(8, 1, .65)$. Pour chaque vecteur π les $k - q$ autres éléments égaient tous 0.05. À des fins de comparaison, nous avons aussi reproduit dans le tableau 5 les résultats relatifs aux cases à probabilités égales.

Tableau 3

Taux d'erreur global¹ pour les intervalles de Bonferroni transformés fondés sur le critère t ;
 $\alpha = 5\%$, $\lambda = 2$, $m = 10$ pour $k \leq 8$, $m = 20$ pour $k = 12$

Taux d'erreur global (ER_T)			
Bonferroni transformés fondés sur le critère t			
k	a	r	Racine carré
Log			
Logit			

¹ Pour $k = 8$ et $r = 50$, la corrélation entre les valeurs estimées de ER_T pour les intervalles log et logit est 0,92. À supposer qu'il s'agit là d'une valeur normale pour tous les r et les k , l'erreur type (selon l'étude de Monte Carlo) de l'écart entre les taux d'erreur des intervalles log et logit est approximativement 0,3 %.

3	.29	15	4.5	4.6	4.6	4.6	4.6
3	.29	30	3.6	4.0	4.0	4.0	4.0
3	.29	50	4.6	5.6	4.7	4.2	3.5
5	.5	15	6.4	4.7	4.5	4.0	3.3
5	.5	30	4.6	4.2	4.5	3.5	3.3
5	.5	50	4.3	4.5	5.9	5.2	4.0
8	.71	15	12.0	5.9	6.6	5.2	4.2
8	.71	30	6.2	6.6	5.4	5.2	4.2
8	.71	50	5.9	5.4	3.9	4.2	4.2
8	.71	100	4.9	3.9	6.7	6.5	6.3
12	.91	15	17.0	6.7	10.1	10.2	10.2
12	.91	30	12.9	6.5	6.3	6.3	6.3
12	.91	50	8.2	6.5	6.3	6.3	6.3

Tableau 4

Degré d'asymétrie ($PER^{(v)}$) du taux d'erreur global pour les méthodes acceptables (en pourcentage);
 $a > 0^2$, $r = 50$, $m = 10$ pour $k \leq 8$, $m = 20$ pour $k = 12$

$PER_U = (ER_U/ER_T) \times 100\%$			
Bonferroni fondés sur le critère t			
α	k	λ	Scheffé (fondés sur le critère F)
Q-H modifiés (premier degré)			
(log)			
(logit)			

5%	5	1.5	19.2	58.7	61.0	48.9
5%	5	2.0	0.0	45.0	61.1	48.8
5%	8	1.5	0.0	63.2	67.5	56.8
5%	8	2.0	0.0	65.2	64.9	49.0
5%	12	2.0	0.0	46.9	53.8	51.6
10%	5	1.5	16.3	49.4	59.2	48.4
10%	5	2.0	6.1	50.0	61.8	48.6
10%	8	1.5	0.0	60.7	67.3	55.8
10%	8	2.0	0.0	65.6	60.7	50.0
10%	12	2.0	0.0	47.5	56.0	51.4

¹ Pour $k = 8$, $\lambda = 2$ et $\alpha = 5\%$, la corrélation entre les valeurs estimées de PER_U pour les intervalles log et logit est 0,82. En supposant qu'il s'agit là d'une valeur normale, les erreurs types (selon l'étude de Monte Carlo) de l'écart entre les degrés d'asymétrie des intervalles log et logit sont approximativement 4 % et 3 % pour $\alpha = 5\%$ et 10 %, respectivement.

² Pour les valeurs de a correspondant à une valeur particulière de k , voir tableau 3.

Il ressort clairement de ce tableau que les deux genres d'ICS sont peu efficaces étant donné le fort taux d'erreur enregistré lorsque le nombre de grappes est relativement peu élevé et que k , le nombre de catégories, est de 5 ou plus. Le fait de recourir à la distribution t de Student pour compenser la variabilité des variances estimées des proportions par catégorie entraîne une nette diminution du taux d'erreur. Toutefois, comme on peut le voir dans le tableau 2, la variante fondée sur la distribution t donne des taux d'erreur encore trop élevés lorsque le nombre de grappes diminue, sauf pour $k = 3$. Si le taux d'erreur tend à augmenter lorsque le nombre de grappes diminue (tant pour les intervalles fondés sur le critère 2 que pour ceux fondés sur le critère 1), c'est à cause de la non-normalité croissante des π_i lorsque r diminue. Cette tendance s'accroît à mesure que k augmente, ce qui est tout à fait naturel étant donné que la non-normalité est de plus en plus prononcée, pour une valeur donnée de r , lorsque les valeurs de π_i diminuent. C'est précisément ce que nous observons dans le cas qui nous occupe, où, rappelons-le, $\pi_i = 1/k \forall i$.

Lorsque $k = 3$, les taux d'erreur pour la variante fondée sur le critère 1 sont essentiellement constants et voisins du seuil nominal. En revanche, pour $k = 8$, ER_T prend des valeurs allant d'environ 8 %, pour $r = 100$, à près de 20 %, pour $r = 15$. D'après les résultats du tableau 2 et d'autres résultats ne paraissant pas ici, il semble que pour $k \geq 8$, les intervalles simples fondés sur le critère 1 tendent très lentement vers leurs limites de Bonferroni lorsque $r \rightarrow \infty$. En outre, pour $k \leq 5$, les taux d'erreurs sont proches du niveau nominal lorsque le nombre de grappes est relativement élevé ($r \geq 40$). Les résultats observés pour $\alpha = 0$ (effets du plan constants) et $\lambda = 1.5$ sont conformes aux résultats ci-dessus. Du point de vue du taux d'erreur global (ou du niveau de confiance), il est clair que c'est seulement si $k \leq 5$ et $r \geq 40$ que l'on peut utiliser les intervalles de Bonferroni simples fondés sur le critère 1 dans des conditions d'échantillonnage en grappes acceptables.

5.5 Intervalles de Bonferroni transformés

Les résultats plus détaillés qui figurent dans Thomas (1989) montrent que l'utilisation de transformations mathématiques ne suffit pas pour résoudre le problème de l'accroissement du taux d'erreur, qui caractérise les intervalles de Bonferroni simples fondés sur le critère 2. En effet, pour les types d'intervalles transformés (racine carrée, log, logit), on observe des taux d'erreur encore très élevés lorsque le nombre de grappes est relativement faible. Heureusement, dans le cas des intervalles fondés sur le critère 1 les transformations ont un tout autre effet, comme l'indiquent les résultats du tableau 3.

Pour $k = 3, 5$ et 8, les taux d'erreur relatifs aux intervalles log et logit sont prêts du niveau nominal de 5 % pour toutes les valeurs r indiquées; ils sont légèrement moins élevés dans le cas des intervalles logit (voir note au bas du tableau 3). En revanche, les intervalles soumis à la transformation racine carrée sont caractérisés par un taux d'erreur excessif lorsque r , est petit et $k \geq 8$; nous ne nous attarderons pas davantage sur ces intervalles. Lorsque le nombre de catégories est élevé ($k = 12$), les intervalles soumis aux transformations log et logit présentent un taux d'erreur assez élevé lorsque le nombre de grappes est moyen ($r = 30$). Cependant, cela ne pose pas vraiment de problème puisqu'en pratique il y a rarement un nombre aussi élevé de catégories. Les résultats observés pour $\alpha = 0$ (effets du plan constants) et $\lambda = 1.5$ sont en général comparables à ceux exposés ci-dessus.

Il semble donc que, pour les valeurs de k, r, λ et α les plus probables en pratique, l'utilisation combinée de transformations log ou logit (qui atténuent la non-normalité dans π) et de valeurs critiques de la distribution t (qui compensent la variabilité des estimations des variances) donne des intervalles ayant le niveau de confiance voulu. Dans la section suivante nous approfondissons l'étude de ces intervalles en analysant la symétrie des taux d'erreur correspondants.

5.3 Intervalles de Quesenberry-Hurst modifiés

Le tableau 1 donne aussi le taux d'erreur global pour les intervalles de Quesenberry-Hurst (Q-H) modifiés du premier degré (section 3.2) lorsque $\alpha = 5\%$, $\lambda = 2$ et $a > 0$.

Le taux d'erreur global est près du niveau nominal de 5% ou nettement au-dessous pour toutes les combinaisons de r et de k indiquées. Lorsque le nombre de grappes est relativement élevé ($r \geq 30$), les taux d'erreur pour $k = 5$, et 8 sont semblables, équivalant à peu près à la moitié du niveau nominal (cela est également vrai pour $k = 12$). Lorsque les effets du plan sont constants (voir Thomas 1989), les intervalles Q-H modifiés du premier degré sont trop étendus pour $k \geq 5$, particulièrement lorsque r est élevé. L'absence de la propriété conservatrice des intervalles de Schэффé pour les intervalles Q-H modifiés dans le cas plus réaliste de l'iné-galité des effets du plan peut s'expliquer à l'aide des arguments de la section 3.1. On voit facilement, d'après l'équation (6), que le niveau de confiance asymptotique des intervalles Q-H modifiés du premier degré est égal à 1 moins la probabilité qu'au moins une des variables aléa-toires $(\hat{\pi}_i - \pi_i)^2 / (\lambda \pi_i (1 - \pi_i) / n)$, $i = 1, \dots, k$, excède asymptotiquement la valeur cri-tique $\chi^2_{k-1}(\alpha)$. Lorsque $a > 0$, ces variables aléatoires ne seront pas toutes distribuées asymptotiquement selon la loi de chi carré avec un degré de liberté, de sorte que la borne défi-nie dans la section 3.1 n'est pas pertinente. La borne qui s'applique réellement au taux d'erreur sera plus élevée puisqu'au moins une des variables aléatoires $(\hat{\pi}_i - \pi_i)^2 / (\lambda \pi_i (1 - \pi_i) / n)$ sera stochastiquement plus grande que $(\hat{\pi}_i - \pi_i)^2 / (v_H / n)$, lorsque $a \geq 0$. On observe des tendances pour $\lambda = 1.5$ (Thomas 1989). En résumé, du point de vue du taux d'erreur global, les intervalles Q-H modifiés du premier degré représentent une méthode de construction d'ICS fiable mais plutôt conservative dans des conditions réalistes d'échantil-lonnage en grappes.

5.4 Intervalles de Bonferroni simples

Le tableau 2 donne le taux d'erreur global pour les intervalles de Bonferroni simples défi-nis par l'équation (9) lorsque $\alpha = 5\%$, $\lambda = 2$, $a > 0$, et $k = 3, 5$ et 8, ainsi que les taux d'erreur correspondants pour la variante de ces intervalles fondée sur le critère t_i qui a été décrite dans la section 3.5.

Tableau 2

Taux d'erreur global pour les intervalles de Bonferroni
simples fondés sur les critères z et t_i ;
 $\alpha = 5\%$, $\lambda = 2$, $m = 10$

Taux d'erreur global (ER _T)				Critère z		Critère t_i	
k	a	r					
3	.29	15	10.0	6.3	5.6	3.29	5.6
3	.29	30	6.3	6.3	4.9	3.29	4.9
3	.29	50	6.5	6.5	5.5	3.29	5.5
5	.50	15	15.0	8.8	9.7	5.50	9.7
5	.50	30	8.8	8.8	7.2	5.50	7.2
5	.50	50	7.2	7.2	5.5	5.50	5.5
8	.71	15	29.6	11.5	19.1	8.71	19.1
8	.71	30	15.0	11.5	11.0	8.71	11.0
8	.71	50	11.5	11.5	9.8	8.71	9.8
8	.71	100	8.1	8.1	7.8	8.71	7.8

compte directement de l'échantillonnage en grappes et qui garantissent un niveau de confiance acceptable pour un nombre de catégories voulu dans diverses conditions d'échantillonnage en grappes.

5.2 Intervalles de Scheffé

Le tableau 1 donne le taux d'erreur global pour les intervalles de Scheffé fondés sur le critère χ^2 (équation (3)) et leur variante fondée sur le critère F , en fonction de r lorsque $\alpha = 5\%$, $\lambda = 2$ et $a > 0$. Des tableaux plus détaillés figurent dans Thomas (1989).

Pour les valeurs de k étudiées, le taux d'erreur global pour les intervalles de Scheffé fondés sur le critère $ER_T \chi^2$ augmente rapidement à mesure que diminue le nombre de grappes, de sorte les intervalles de ce genre ne devraient jamais être utilisés pour un petit nombre de grappes. Par contre, la variante de cet intervalle fondée sur le critère F affiche un taux d'erreur global qui se maintient autour de $\alpha = 5\%$ dans tous les cas. Lorsque r augmente, le taux d'erreur global pour la variante fondée sur le critère F demeure à peu près le même pour $k = 3$ mais diminue rapidement pour $k \geq 5$, on observe la même tendance pour l'intervalle fondé sur le critère χ^2 . Ces tendances empiriques sont le résultat de deux phénomènes concomitants. Lorsque r augmente, les taux d'erreur pour les deux genres d'intervalles tendent vers leurs niveaux asymptotiques, qui ont pour bornes supérieures 4.29, 1.04 et 0.14 % pour $k = 3, 5$ et 8 respectivement (voir section 3.1). Toutefois, lorsque r diminue, le conservatisme des intervalles de Scheffé (pour $k \geq 5$) est graduellement affaibli par les effets de la non-normalité croissante des proportions estimées, π . L'augmentation du taux d'erreur attribuable à la non-normalité est moins prononcée dans le cas de la variante fondée sur le critère F que dans le cas de l'intervalle fondé sur le critère chi carré (équation 3), de sorte que ER_T ne dépasse jamais de beaucoup le niveau nominal de 5 % dans le premier cas. Pour un effet de grappes modéré ($\lambda = 1.5$), la variante fondée sur le critère F produit des résultats qualitativement comparables à ceux observés pour $\lambda = 2$. Par conséquent, du point de vue du taux d'erreur global, l'intervalle de Scheffé fondé sur le critère F peut être utilisé dans de nombreuses conditions d'échantillonnage en grappes mais son conservatisme représente un inconvénient.

Tableau 1
Taux d'erreur global pour les intervalles de Scheffé et les intervalles Q-H modifiés;
 $\alpha = 5\%$, $\lambda = 2$, $m = 10$

Taux d'erreur global (ER_T)							
Scheffé (fondés sur le critère χ^2)	Scheffé (fondés sur le critère F)	Q-H modifiés (premier degré)	k	a	r		
9.2	5.9	5.0	3	.29	15	3	3
5.7	4.7	5.1	3	.29	30	3	3
5.4	5.0	5.4	3	.29	50	3	3
8.8	5.2	4.3	5	.5	15	5	5
4.0	3.0	2.7	5	.5	30	5	5
2.5	2.0	2.0	5	.5	50	5	5
12.7	7.4	2.4	8	.71	15	8	8
4.2	3.0	2.7	8	.71	30	8	8
2.7	1.6	2.5	8	.71	50	8	8
0.8	0.7	2.3	8	.71	100	8	8

(ER_i) pour les besoins de cet article, est substitué en l'occurrence au niveau de confiance réel (1 — taux d'erreur global) car il peut être décomposé facilement en deux taux qui permettront de juger de la symétrie ou de la "propriété d'être sans biais" de chaque méthode de construction d'ICS. Jennings (1987) soutient que le fait de considérer uniquement les niveaux de confiance peut entraîner une fausse évaluation des méthodes de construction d'intervalles de confiance pour un paramètre et recommande de prendre en considération le nombre de fois qu'un intervalle se situe à droite ou à gauche de la valeur réelle du paramètre. Dans cet article, nous avons appliqué la recommandation de Jennings à des intervalles de confiance simultanés pour $\pi_i, i \in I$, où I est l'ensemble d'indices $\{1, \dots, k\}$, en dénombrant les essais de Monte Carlo où :

- (a) il y avait un plus grand nombre d'intervalles à droite qu'à gauche de $\pi_i, i \in I$;
- (b) il y avait un plus grand nombre d'intervalles à gauche qu'à droite de $\pi_i, i \in I$;
- (c) il y avait autant d'intervalles (> 0) à droite qu'à gauche de $\pi_i, i \in I$.

Les taux d'erreur supérieur et inférieur sont alors définis $ER_U = [n_a + (n_c/2)]/N_i$ et $ER_L = [n_b + (n_c/2)]/N_i$ respectivement, où N_i représente le nombre d'essais de Monte Carlo et n_a, n_b et n_c sont les chiffres correspondant à a), b) et c) respectivement. La somme de ER_U et de ER_L est manifestement égale au taux d'erreur global, ER_T . Les taux d'erreur inférieur et supérieur serviront à comparer des méthodes de construction d'ICS dont le taux d'erreur global est suffisamment près du niveau nominal α , en fonction d'une série de valeurs de paramètres et de divers scénarios d'échantillonnage en grappes. Nous avons aussi calculé des longueurs d'intervalles moyennes et les erreurs types correspondantes; ces valeurs serviront de critère final pour le choix des meilleures méthodes.

5. SOMMAIRE DES RÉSULTATS DE L'ÉTUDE DE MONTE CARLO

Tous les résultats présentés dans cette section concernent le taux d'erreur global ER_T , qui a été défini dans la section 4. Faute d'espace, les tableaux de cette section ne concernent que le cas où les effets du plan sont inégaux ($\alpha > 0$) et $\lambda = 2$. Le lecteur trouvera des résultats plus détaillés dans Thomas (1989). Aux fins de l'interprétation des résultats présentés dans ces tableaux, soulignons que pour 1,000 essais de Monte Carlo, les erreurs types binomiales des valeurs estimées de taux d'erreur globaux de 5, 10 et 20 % sont 0,7, 0,9 et 1,3 % respectivement. En règle générale, nous ne considérerons que les écarts (écart par rapport au seuil nominal ou écart entre les taux d'erreur de diverses méthodes de construction d'ICS) qui seront suffisamment grands pour avoir une signification pratique et qui équivaldront au moins au double de l'erreur type de Monte Carlo correspondante.

5.1 Intervalles multinomiaux

Nous ne donnons ici qu'un résumé des résultats concernant les intervalles multinomiaux; pour plus de détails, le lecteur est prié de se référer à Thomas (1989). Suivant un échantillonnage en grappes, les taux d'erreur pour les intervalles de Bonferroni de Goodman (voir équation (9), où p_{ij} est remplacé par $\pi_i (1 - \pi_i)$) atteignent un niveau inacceptable sauf lorsque λ est proche de 1, c.-à-d. lorsque l'effet de grappe est faible. En revanche, les intervalles de Scheffé-Gold et de Quesenberry-Hurst ont parfois des taux d'erreur qui se rapprochent du niveau nominal, plus particulièrement lorsque leur conservatisme compense les effets de grappe, qui tendent normalement à accroître le taux d'erreur (voir aussi Andrews et Birdsall 1988). Malheureusement, cela n'est pas toujours le cas; ces deux genres d'intervalles affichent parfois des taux d'erreur excessifs ($ER_T \geq 2\alpha$) pour des combinaisons réalistes du nombre de catégories et du modèle d'échantillonnage en grappes. Par conséquent, les intervalles multinomiaux ne devraient pas être utilisés pour des données d'enquêtes complexes. De toute évidence, il est nécessaire d'élaborer des méthodes qui tiennent

4. PLAN DE L'ÉTUDE DE MONTE CARLO

4.1 Paramètres et nombres aléatoires

Les paramètres considérés sont : i) le niveau de confiance nominal $(1 - \alpha)$ de l'ICS; ii) π , le vecteur de probabilités du modèle; iii) k , le nombre de catégories; iv) r , le nombre de grappes échantillonnées; v) m , le nombre d'unités prélevées dans chaque grappe échantillonnée; vi) λ , la moyenne des effets du plan généralisés (valeurs propres); vii) a , le coefficient de variation des effets du plan généralisés. La nature et le degré de l'échantillonnage en grappes sont représentés par la paire de valeurs (λ, a) selon les termes suivants : a) échantillonnage multinomial ($\lambda = 1, a = 0$); b) échantillonnage avec effets du plan constants ($\lambda > 1, a = 0$); c) échantillonnage avec effets du plan non constants ($\lambda > 1, a > 0$).

Des simulations de Monte Carlo ont été effectuées pour des combinaisons particulières de k, λ, a et r_{max} , qui est le nombre maximum de grappes que l'on peut obtenir dans une phase de traitement. La plupart des simulations ont été exécutées avec deux valeurs de λ , soit 1.5 et 2.0, et deux valeurs de a , soit $a = 0$ (effets du plan constants) et $a > 0$ (effets du plan non constants), pour des catégories équiprobables ($\pi_i = 1/k, i = 1, \dots, k$). Au départ, nous avons choisi trois valeurs de k ($k = 3, 5, 8$) qui reflètent le nombre de catégories que l'on trouve le plus souvent dans les tests de validité de l'ajustement. Par la suite, nous avons procédé à une simulation avec $k = 12, \lambda = 2$ et $a > 0$ dans le but de vérifier le champ d'applicabilité des résultats. Le nombre d'unités par grappe a été fixé à $m = 10$ pour $k = 3, 5$ et 8, et à $m = 20$ pour $k = 12$. Des analyses préliminaires ont montré que ce paramètre n'avait aucun effet sur les niveaux de confiance. Pour pouvoir comparer les résultats en fonction de k , nous avons déterminé les valeurs non nulles de a de manière que a/a^{max} soit identique pour toutes les valeurs de k , choisies, $a^{max} = (k - 2)^{1/2}$ étant la valeur maximum que peut prendre a . Pour $k = 5$, nous avons fixé la valeur de a à 0.5, qui est la valeur que l'on retrouve le plus souvent en pratique (p. ex. $a = 0.43$ pour $k = 5$, selon Rao et Thomas 1988).

Le fait de concentrer initialement notre attention sur des catégories équiprobables nous a permis de faire une analyse efficace de l'incidence de k , de λ et a sur les niveaux de confiance et d'exclure de l'analyse une bonne partie des variantes d'ICS qui peuvent exister. Les simulations additionnelles dont nous faisons état à la section 7 montrent que les méthodes qui ont passé cette étape peuvent, de fait, être appliquées lorsque les probabilités par case diffèrent sensiblement les unes des autres. Les vecteurs de probabilités inégales ont été limités à la classe $\pi(k, q, \phi)$, définie par les éléments $\pi_i = \phi, i = 1, \dots, q$ et $\pi_i = (1 - q\phi)/(k - q), i = q + 1, \dots, k$.

Pour avoir plus de détails sur la production des grappes aléatoires à partir de la distribution multinomiale de Dirichlet, le lecteur est prié de consulter Thomas et Rao (1987). Chaque simulation de Monte Carlo consistait en 1,000 séries de grappes, chacune pouvant contenir jusqu'à 100 grappes indépendantes formant des sous-ensembles emboîtés. La simulation a appliqué tout à tour les méthodes de construction d'ICS à chaque sous-ensemble en utilisant deux niveaux de confiance nominaux (95 et 90 %); on a ainsi pu faire des comparaisons plus précises, premièrement entre les résultats obtenus avec une même méthode en faisant varier le nombre de grappes. La plupart des résultats présentés se rapportent au niveau de 95 %, ceux relatifs au niveau de 90 % étant, dans l'ensemble, qualitativement comparables aux premiers.

4.2 Méthodes d'évaluation

On procède à un tri préliminaire des principales méthodes de construction d'ICS en se fondant sur le pourcentage d'essais de Monte Carlo où au moins un des k intervalles de confiance ne renferme pas la valeur réelle du paramètre. Ce pourcentage sert à mesurer le taux d'erreur de famille, qui équivaut au seuil de signification réel de l'ICS lorsque celui-ci sert au test de validité de l'ajustement. Le taux d'erreur de famille, que nous appellerons taux d'erreur global

3.3 Intervalles de Bonferroni simples

Comme grosso modo, il est permis d'affirmer que chaque π_i est distribuée asymptotique-ment selon une loi $N(\pi_i, v_{ii}^H/n)$, les intervalles

$$(9) \quad \pi_i \in \left\{ \pi_i \pm (v_{ii}^H/n)^{1/2} z_{\alpha'/2} \right\},$$

où $\alpha' = \alpha/k$ et $z_{\alpha'/2}$ est la limite supérieure ($\alpha'/2$ pour cent) de la distribution normale cen-trée réduite, auront un niveau de confiance d'au moins $(1 - \alpha)$ pour les grands échantillons grâce à l'inégalité de Bonferroni. Ils sont équivalents aux intervalles de Scheffé, la variable A de l'équation (3) étant remplacée par $A^{(3)} = \chi^2_1(\alpha')$. Comme le souligne Goodman (1965), ils seront moins étendus que les intervalles de Scheffé pour les valeurs courantes de α et de k (p. ex.: $\alpha = 1\%$, 5% , ou 10%). Les intervalles de Bonferroni multinomiaux de Goodman (1965) sont aussi définis par l'équation (9) sauf que v_{ii}^H est remplacé par $\pi_i(1 - \pi_i)$. Les extrê-mités des intervalles de Bonferroni simples qui se situent à l'extérieur de $[0, 1]$ seront rame-nées à 0 ou 1 selon le cas.

3.4 Intervalles de Bonferroni transformés

Pour une fonction g suffisamment lisse, $g(\hat{\pi})$ sera distribuée asymptotiquement selon une loi $N(g(\pi_i), [g'(\pi_i)]^2 v_{ii}^H/n)$, où $g'(\pi_i)$ désigne la dérivée partielle $\partial g(\pi_i)/\partial \pi_i$ évaluée à π_i . On peut alors déterminer les intervalles de Bonferroni en prenant l'inverse de la fonction des inter-valles correspondants pour les $g(\pi_i)$, ce qui donne

$$(10) \quad \pi_i \in \left\{ g^{-1}(g(\hat{\pi})) \pm g'_i(\hat{\pi})(v_{ii}^H/n)^{1/2} z_{\alpha'/2} \right\}.$$

Trois fonctions g seront étudiées: la racine carrée $g_1(\pi_i) = \pi_i^{1/2}$ (qui a déjà été analysée par Bailey, 1980, pour l'échantillonnage multinomial), le logarithme naturel $g_2(\pi_i) = \ln(\pi_i)$; et le logit $g_3(\pi_i) = \ln(\pi_i)/(1 - \pi_i)$. Les extrémités des intervalles qui se situent à l'extérieur de $[0, 1]$ seront une fois de plus ramenées à 0 ou à 1 selon le cas. Par ailleurs, nous nous sommes intéressés aux intervalles de Bonferroni transformés qui reposent sur un estimateur de la variance de $g(\hat{\pi})$ déterminé à l'aide d'une méthode jackknife (voir Thomas 1989). Nous avons pu constater que l'utilisation d'un tel estimateur présentait peu d'intérêt; il est donc recommandé d'utiliser les estimateurs déterminés à l'aide de la série de Taylor en raison de leur simplicité. Nous ne nous attarderons pas davantage sur les inter-valles qui reposent sur des estimateurs de variance déterminés par "jackknife".

3.5 Variantes des intervalles définis ci-dessus

Intervalles de Scheffé: D'après Thomas et Rao (1987), on peut modifier les intervalles de Scheffé en remplaçant la valeur critique A dans l'équation (3) par $A^{(4)} = (k - 1)(r - 1)(r - k + 1)^{-1} F_{(k-1), (r-k+1), (r-k+1), (r-k+1)}(\alpha)$, où $F_{(k-1), (r-k+1), (r-k+1), (r-k+1)}(\alpha)$ est la borne supérieure (α pour cent) d'une distribution F avec $(k - 1)$ et $(r - k + 1)$ degrés de liberté. **Intervalles de Quesenberry-Hurst:** On peut aussi définir des variantes des intervalles de Quesenberry-Hurst ($Q-H$) modifiés; il s'agit en l'occurrence d'une version F' d'une variable à tester soumise à des corrections du premier et du second degré comme le proposent Thomas et Rao (1987). Les intervalles correspondants sont, là encore, plus étendus qu'il ne le faut et ne seront pas considérés dans cet article.

Intervalles de Bonferroni: Des arguments heuristiques (voir l'annexe de Thomas et Rao 1987) donnent à penser que l'on peut améliorer les intervalles de Bonferroni simples en remplaçant $z_{\alpha'/2}$ par $t_{r-1}(\alpha'/2)$, qui est la valeur critique supérieure ($\alpha'/2$ pour cent) d'une distri-bution t de Student avec $r - 1$ degrés de liberté. Cette méthode vaut aussi pour les intervalles de Bonferroni transformés.

certain nombre des ℓ directions possibles (voir Miller 1981, p. 63). De fait, comme il est possible de le montrer à l'aide de l'argument de Goodman (1965) présenté ci-dessous, ces intervalles seront de plus en plus étendus à mesure que k augmentera. Le niveau de confiance des intervalles de Scheffé est égal à 1 moins la probabilité que se produise au moins un des événements $\{(\pi_i)^2/(\hat{v}_{H}/n) > \chi_{2(k-1)}^2(\alpha)\}$, $i = 1, \dots, K$; comme les variables aléatoires $(\pi_i - \pi_i)^2/(\hat{v}_{H}/n)$ sont toutes distribuées asymptotiquement selon une loi de chi-carré avec un degré de liberté, il est possible d'évaluer la probabilité de chacun des événements. Par l'indépendance de Bonferroni, on peut déterminer des bornes inférieures pour un niveau de confiance donné. Ainsi, pour un niveau nominal de 95 % et des valeurs $k = 3, 5, 8$ et 12, ces bornes sont .9571, .9896, .9986 et .9999 respectivement.

3.2 Intervalles de Quesenberry-Hurst modifiés

Suivant l'hypothèse d'un échantillonnage multinomial, Quesenberry et Hurst (1964) ont résolu la proposition probabiliste pour grands échantillons

(5)
$$P\left\{X^2 = n \sum_{i=1}^k \frac{\pi_i}{(\pi_i - \pi_i)^2} \leq A\right\} = 1 - \alpha$$

pour les probabilités par case π_i , et ont obtenu les ICS

(6)
$$\pi_i \in \left\{ \frac{\pi_i + A/2n \pm (A/n)^{1/2} [\pi_i (1 - \pi_i) + A/4n]^{1/2}}{1 + A/2n} \right\}.$$

Selon un échantillonnage multinomial, ces intervalles sont asymptotiquement équivalents à ceux de Scheffé et de Scheffé-Gold et ont donc les mêmes caractéristiques que ceux-ci au point de vue du niveau de confiance. Il est possible d'utiliser les intervalles de Quesenberry-Hurst (Q-H) pour des données d'échantillon en grappes en appliquant, comme le proposent Rao et Scott (1981), les corrections du premier et du second degré à la distribution de X^2 . On obtient les ICS du premier et du second degré correspondants en remplaçant A dans l'équation (3) par

(7)
$$A^{(1)} = \hat{\lambda} A \quad \text{et} \quad A^{(2)} = \hat{\lambda} (1 + a^2) \chi^2_v(\alpha)$$

respectivement, où $v = (k - 1)/(1 + a^2)$ et $\hat{\lambda}$, un estimateur de la moyenne des effets du plan généralisés, est défini (Rao et Scott 1981)

(8)
$$\hat{\lambda} = (k - 1)^{-1} \sum_{i=1}^k (1 - \pi_i) d_i,$$

où $d_i, i = \dots, k$ est la valeur estimée de l'effet du plan par case et est déterminée par l'expression $d_i = \hat{v}_{H}/\pi_i (1 - \pi_i)$. On calcule la valeur estimée du coefficient de variation, a , en remplaçant $\hat{\lambda}$ dans l'équation (1) par $\hat{\lambda}$ et $\sum \lambda_i^2$ par la valeur estimée $\sum \hat{\lambda}_i^2 = \sum \hat{v}_{H}^2/\pi_i \pi_i$. Il ressort de l'analyse de Thomas (1989) que les intervalles modifiés du second degré sont plus étendus qu'il ne le faudrait; par conséquent, nous ne nous intéresserons qu'aux intervalles Q-H du premier degré.

m unités du second degré sont tirées d'une distribution multinomiale dans le cas de chaque grappe, selon la valeur réalisée de p_i pour la grappe. Désignons le vecteur d'effectifs pour chaque grappe par $m_i = (m_{i1}, \dots, m_{ik-1})$, ou $m_k = m - \sum_{i=1}^{k-1} m_i$. Ainsi pour l'échantillon tout entier, $m = \sum_{i=1}^k m_i$ et en termes de proportions, $\hat{\pi} = \sum_{i=1}^k \hat{\pi}_i$ où $\hat{\pi}_i = m_i/m$. Briër (1981) a montré que, selon ce modèle, $E(\hat{\pi}) = \pi$ et $V(\hat{\pi}) = dP/n$, en d'autres termes, la matrice des covariances de $\hat{\pi}$ est proportionnelle à la matrice des covariances multinomiale, avec la constante de proportionnalité de $d > 1$. Selon ce modèle, la matrice des effets du plan est définie $D = dI^{k-1}$, où I^{k-1} est la matrice unité d'ordre $k - 1$. Ainsi, $\lambda_i = d \forall i$, de sorte que $\lambda = d$ et $a = 0$. Le modèle de Briër peut donc représenter l'augmentation de variance ($\lambda > 1$), mais non les effets du plan généralisés inégaux que l'on observe dans la pratique. Thomas et Rao (1987) ont utilisé une variante du modèle de Briër, selon laquelle les vecteurs p_i sont tirés de façon indépendante d'une combinaison de deux distributions de Dirichlet, qui rend l'idée d'une population composée de deux catégories de grappes distinctes. Ce modèle, qui est un cas particulier du modèle proposé par Rao et Scott (1979), produit $k - 2$ valeurs propres identiques et une valeur propre distincte, λ et a étant des fonctions explicites des paramètres de Dirichlet. L'étude de Monte Carlo s'en trouve largement simplifiée car on peut contrôler avec satisfaction les valeurs de λ et de a . Comme le modèle de Thomas et Rao (1987) respecte les conditions fondamentales énoncées plus haut ($\lambda > 1, a > 0$), c'est lui que nous utiliserons.

3. INTERVALLES DE CONFIANCE SIMULTANÉS

3.1 Intervalles de Scheffé

Un argument courant de Scheffé fondé sur le jugement de probabilité asymptotiquement exact

$$(2) \quad P\left(n(\hat{\pi} - \pi)' V^{-1}(\hat{\pi} - \pi) \leq \chi^2_{k-1}(\alpha)\right) = 1 - \alpha$$

permet d'établir des intervalles de confiance simultanés pour des combinaisons linéaires, $\ell' \pi$, des probabilités par catégorie, où ℓ est un vecteur de dimension ($k - 1$). Des valeurs de ℓ appropriées permettent de construire des ICS pour les probabilités par case, ces intervalles étant définis

$$(3) \quad \pi_i \in \left\{ \hat{\pi}_i \pm (\hat{v}_i^H)^{1/2} (A/n)^{1/2} \right\}, i = 1, \dots, k,$$

où $A = \chi^2_{k-1}(\alpha)$ est la limite supérieure (à α pour cent) d'une distribution chi-carré avec $k - 1$ degrés de liberté, et \hat{v}_i^H est l'élément diagonal i d'un estimateur convergent de V (lorsque $r \rightarrow \infty$) défini

$$(4) \quad V = \frac{n}{r(r-1)} \sum_{i=1}^r (\hat{\pi}_i - \hat{\pi})(\hat{\pi}_i - \hat{\pi})'$$

Notons que lorsque l'extrémité d'un intervalle se situe à l'extérieur de $[0, 1]$, il faut modifier la définition (3) en ramenant l'extrémité vers 0 ou 1 selon le cas. En ce qui concerne l'échantillonage multinomial, on peut remplacer \hat{v}_i^H par $\hat{\pi}_i(1 - \hat{\pi}_i)$, auquel cas les intervalles de Scheffé se ramènent à ceux proposés par Gold (1963). Nous parlerons alors des intervalles de Scheffé-Gold. Les intervalles de Scheffé définis en (3) seront prudents, c'est-à-dire qu'ils auront un seuil de confiance qui excédera asymptotiquement ($1 - \alpha$) puisqu'ils n'utilisent qu'un

L'estimation des intervalles de confiance simultanés (ICS) est un complément important des tests d'hypothèses. Cet article est donc la suite logique de l'analyse de Thomas et Rao (1987) portant sur les critères utilisés dans les tests de validité de l'ajustement selon un échantillonnage en grappes simulé. Dans cet article, nous proposons des modifications aux méthodes courantes de construction d'ICS et évaluons l'efficacité des nouvelles versions pour de petits échantillons au moyen d'une étude de Monte Carlo.

Dans la section 2, nous décrivons le modèle d'échantillonnage en grappes utilisé dans l'étude de Monte Carlo tandis que dans la section 3, nous présentons les méthodes de construction d'ICS qui seront analysées. Dans la section 4, nous exposons le plan de la simulation de Monte Carlo ainsi que les méthodes servant à évaluer l'efficacité des intervalles de confiance. Enfin, les sections 5, 6 et 7 contiennent les principaux résultats de l'étude et la section 8 renferme les conclusions ainsi que des recommandations.

2. LE MODÈLE D'ÉCHANTILLONNAGE EN GRAPPES

Notre étude porte plus spécialement sur un plan d'échantillonnage à deux degrés, selon lequel un échantillon de m unités à k catégories est prélevé de façon indépendante dans chacune des r grappes préalablement échantillonnées.

Pour un échantillon de taille $n = mr$, définissons $m = (m_1, \dots, m_{k-1})'$ comme le vecteur des effectifs par catégorie pour tout l'échantillon, où $m_k = n - \sum_{i=1}^{k-1} m_i$. Par ailleurs, définissons $\pi = (\pi_1, \dots, \pi_{k-1})' = m/n$ comme le vecteur des proportions par catégories pour tout l'échantillon. En outre, posons $\pi = E(\tilde{\pi})$, où E désigne l'espérance mathématique suivant un modèle d'échantillonnage en grappes approprié, et définissons V/n comme la matrice $(k-1) \times (k-1)$ des covariances de $\tilde{\pi}$. D'après Rao et Scott (1981), l'effet du plan ordinaire pour la combinaison linéaire $c'\tilde{\pi}$ des proportions par catégorie est $c'Vc/c'Pc$, où P équivaut à n fois la matrice des covariances de $\tilde{\pi}$ suivant un échantillonnage multinomial, c.-à-d. $P = \text{diag}(\pi) - \pi\pi'$, et c est un vecteur de dimension $k-1$. L'effet du plan maximum calculé pour toutes les combinaisons linéaires possibles correspond à la valeur propre la plus élevée de la matrice des effets du plan $D = P^{-1}V$. Les valeurs propres de D , identifiées en ordre décroissant par $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$, sont désignées par Rao et Scott (1981) comme des effets du plan généralisés et constituent un sommaire quantitatif de l'augmentation de variance liée à un plan d'échantillonnage particulier par rapport à l'échantillonnage aléatoire simple. Suivant la distribution multinomiale, qui correspond à un échantillonnage aléatoire simple dans de grandes populations, $\lambda_j = 1 \forall j$. Les plans qui comportent un échantillonnage en grappes produisent habituellement des effets du plan généralisés qui sont en moyenne supérieurs à 1, c.-à-d. $\lambda = \sum_{j=1}^{k-1} \lambda_j / (k-1) > 1$. De plus, des analyses de données d'enquête réelles (Hidiroglou et Rao 1987; Rao et Thomas, 1988) ont révélé une forte variabilité des λ_j . Cette variabilité est représentée concrètement par le coefficient de variation des λ_j , qui est défini

$$a = \left(\sum_{j=1}^{k-1} \lambda_j^2 / [(k-1)\lambda^2] - 1 \right)^{1/2}. \quad (1)$$

Un modèle d'échantillonnage en grappes acceptable doit donc pouvoir produire des effets du plan généralisés tels que $\lambda > 1$ et $a > 0$.

Brier (1981) a proposé un modèle d'échantillonnage en grappes à deux degrés, où chaque grappe est représentée par un vecteur de probabilités par catégorie, $p_i = (p_{i1}, p_{i2}, \dots, p_{i, k-1})'$, $i = 1, \dots, r$, où $p_{ik} = 1 - \sum_{i=1}^{k-1} p_{it}$. Chaque vecteur p_i est tiré de façon indépendante d'une distribution de Dirichlet avec une moyenne π , c.-à-d. $E(p_i) = \pi$, tandis que les

Intervallles de confiance simultanéés pour proportions suivant un modèle d'échantillonnage en grappes

D. ROLAND THOMAS¹

RÉSUMÉ

L'auteur analyse par une étude de Monte Carlo des méthodes de construction d'intervallles de confiance simultanéés pour $k > 2$ proportions selon un modèle d'échantillonnage en grappes à deux degrés. Parmi les intervallles de confiance étudiés, citons i) les intervallles multinomiaux ordinaires, ii) les intervallles de Scheffé fondés sur des estimations-échantillon des variances de proportions de case, iii) les intervallles de Quesenberry-Hurst adaptés à des données agglomérées au moyen des corrections de premier et de second degré de X^2 de Rao et Scott, iv) les intervallles de Bonferroni simples, v) les intervallles de Bonferroni fondés sur des transformations des proportions estimées, et vi) les intervallles de Bonferroni calculés au moyen des niveaux critiques du test t de Student. L'étude de Monte Carlo révèle que, dans plusieurs situations, le niveau de confiance réel des intervallles multinomiaux est largement inférieur au niveau théorique. Les intervallles les plus efficaces au point de vue du niveau de confiance et de la symétrie des taux d'erreur (notion découlant d'un principe avancé par Jennings) sont les intervallles de Bonferroni fondés sur le critère t et soumis aux transformations logarithmique et logit. Parmi les intervallles de type Scheffé, les plus efficaces sont les intervallles de Quesenberry-Hurst modifiés par la correction de premier degré de Rao-Scott.

MOTS CLÉS : Inférence simultanéée; enquêtes complexes; Monte Carlo.

1. INTRODUCTION

Les résultats d'enquête prennent souvent la forme de proportions ou de pourcentages estimés des unités de population qui appartiennent à plus d'une catégorie. C'est le cas, notamment, des résultats de nombreuses études sociologiques (voir, par exemple, Black et Myles 1986), des études de mise en marché et des sondages d'opinion. Comme le soulignent Fitzpatrick et Scott (1987), l'inférence portant sur les proportions par catégorie repose souvent sur des intervallles de confiance binomiaux même lorsque l'analyse porte sur plus de deux catégories. Dans cet article, nous analysons plusieurs méthodes de construction d'intervallles de confiance simultanéés pour les proportions $\pi_i, i = 1, \dots, k$, des unités de populations qui appartiennent à l'une ou l'autre de k catégories distinctes et ce, à l'aide des données d'un échantillon en grappes à deux degrés. Les méthodes courantes de construction d'intervallles de confiance simultanéés pour données qualitatives, qui ont fait l'objet d'une analyse par Hochberg et Tamane (1987), reposent sur l'hypothèse de la distribution multinomiale des données d'échantillon et sont donc indiquées pour des données d'échantillonnage aléatoires simples. Lorsqu'il s'agit de données recueillies au moyen d'un plan d'échantillonnage en grappes, les méthodes courantes sont vraisemblablement peu efficaces; on constate la même faiblesse lorsque des tests fondés sur une distribution multinomiale sont appliqués à des données d'enquêtes complexes. À ce propos, de nombreux auteurs ont montré que l'échantillonnage en grappes pouvait engendrer des erreurs de première espèce excessivement élevées (voir, par exemple, Fellegi 1980, Rao et Scott 1979 et 1981, Holt, Scott et Ewing 1980). Ainsi, en ce qui concerne les intervallles de confiance simultanéés, nous devons naturellement nous attendre que l'échantillonnage en grappes implique que des probabilités de couverture inférieures à ce qu'indique la théorie multinomiale.

¹ D. Roland Thomas, School of Business, Carleton University, Ottawa, Ontario, K1S 5B6.

lorsque la matrice des covariances des estimations par case, V , est proportionnelle à la matrice des covariances multinomiale, F . Cependant, le test F est moins puissant que les tests de Rao-Scott, à moins que le nombre de degrés de liberté au dénominateur soit élevé. Dans un tel cas, le test F pourrait être efficace même si la condition $V \propto F$ n'est pas satisfaite (voir Rao et Scott 1987, p. 392).

En ce qui concerne les données du tableau 1, $F = 6.63$ pour le test de l'hypothèse $\gamma = 0$ étant donné le modèle (2.12); ce test n'est pas significatif à un seuil de 5 % puisque $F_{1,3}(0.05) = 10.01$, (limite supérieure de la distribution F à 1 % avec 1 et 3 degrés de liberté (d.l.)). Par contre, le test de Wald W_1 et le test de Rao-Scott, qui exigent tous deux des renseignements détaillés sur la matrice des covariances estimée, sont significatifs à un seuil de 1 % puisque $\chi^2_1(0.01) = 6.63$. Le test F semble donc moins puissant en l'occurrence puisque le nombre de d.l. au dénominateur n'est que de 3. Le test que propose Molina est, de fait, l'équivalent d'un test F mais Molina tient la variable F pour une variable χ^2 avec un d.l., ce qui n'est peut-être pas juste étant donné le faible nombre de d.l. au dénominateur.

Avec la méthode GLIM, on ne peut obtenir de critère pour tester la validité de l'ajustement d'un modèle. Pour qu'un test de validité de l'ajustement soit conforme, il faut disposer de renseignements sur les effets du plan.

(iii) Réponse aux commentaires de C.J. Skinner

Skinner a précisé que l'on pourrait définir le test d'équivalence de deux modèles de régression logistique décrit dans la section 4 comme un test d'hypothèse emboîtée selon le modèle de Roberts, Rao et Kumar (1987) en utilisant des variables x auxiliaires. Or, le modèle de Roberts, Rao et Kumar suppose un échantillon de taille fixe n alors que dans la section 4, il y a deux échantillons de tailles fixes respectives n_1 et n_2 pour les deux périodes. Il faudrait donc modifier soit généralement les résultats obtenus par Roberts, Rao et Kumar pour pouvoir les appliquer au test d'équivalence de deux modèles de régression logistique. En outre, l'utilisation de variables auxiliaires implique qu'il faut estimer de façon itérative la valeur de 2s paramètres tandis que dans la section 4, il est question de deux solutions itératives, chacune comprenant seulement s paramètres. L'utilisation de variables auxiliaires pourrait donc créer des problèmes de convergence si s est grand.

Les MCP avec matrices de covariances singulières ont été étudiés exclusivement dans la section 2 puisque les modèles logit étudiés dans les autres sections ne comprenaient pas de matrice des covariances singulière. La méthode des MCP peut aussi s'appliquer à des modèles logit mais les estimateurs que l'on obtient et les tests de Wald correspondants peuvent être instables si le C3 de Skinner). Les six critères que propose Skinner pour comparer les méthodes MCP et PMV sont très utiles. En ce qui concerne l'efficacité relative des estimateurs MCP et PMV selon des plans de sondage complexes, il n'existe pas de résultats généraux mais les estimateurs MCP ne devraient pas être beaucoup plus efficaces (en fait, ils pourraient être moins efficaces) si le nombre de degrés de liberté rattaché à la matrice des covariances estimée est faible. De toute évidence, il serait utile de poursuivre la recherche sur l'efficacité relative des estimateurs MCP et PMV.

SOURCES ADDITIONNELLES

RAO, J.N.K., et COLIN, D. (1988). Fitting dose-response models and hypothesis testing in teratological studies. Rapport technique N° 116, Laboratory for Research in Statistics and Probability, Carleton University et l'Université d'Ottawa, Ottawa, Ontario.

RÉPONSE DES AUTEURS

Nous remercions les trois participants, MM. Fay, Molina et Skinner, de nous avoir communiqué leurs précieux commentaires et de nous avoir proposé d'autres méthodes pour l'analyse de données recoupées provenant d'enquêtes complexes.

(i) Réponse aux commentaires de R.E. Fay

Nous reconnaissons avec Fay que la méthode de la répétition et les tests chi carré avec estimateur jackknife peuvent remplacer adéquatement les méthodes présentées dans notre article, à la condition que le plan de sondage permette l'utilisation de méthodes de répétition comme le "jackknife" ou la répétition compensée. De fait, le programme CPLX que propose Fay offre une méthode d'analyse complète dans les cas où on dispose d'estimations pour chaque échantillon répété. De plus, comme nous le soulignons dans l'introduction, des études de simulation montrent que les tests "jackknife" de Fay et les corrections de Rao-Scott donnent des résultats acceptables dans des conditions générales, ce qui n'est pas le cas des tests de Wald fondés sur les moindres carrés pondérés. Toutefois, les corrections de Rao-Scott peuvent aussi être appliquées à des plans de sondage qui ne permettent pas l'utilisation d'une méthode de répétition. Les progiciels conçus pour l'Enquête Santé Canada et l'enquête sur la population active du Canada permettent de calculer directement la matrice des covariances estimée des estimations par case mais non celle des estimations d'échantillon répété. Par conséquent, il faudrait modifier les progiciels avant d'appliquer des tests "jackknife".

À notre grande satisfaction, Fay souligne que les méthodes exposées dans notre article et les méthodes analogues découlant de la théorie de la répétition peuvent aussi servir à résoudre des problèmes d'inférence pour des expériences à plan de sondage complexe, où il est question d'échantillonnage en grappes et de stratification. En effet, l'un de nous (J.N.K. Rao) a utilisé récemment des méthodes du type Rao-Scott pour ajuster des modèles dose-réponse et tester des hypothèses dans des études téralogiques où les unités expérimentales étaient des portées d'animaux (Rao et Colin 1989). Contrairement à d'autres méthodes proposées dans ce domaine, les méthodes utilisées par Rao et Colin ne supposent pas des modèles précis pour les corrélations intraportée. Nous avons considéré les modèles de transformation de Box-Cox puisque Guettero et Johnson (1982) ont obtenu avec ces modèles des ajustements beaucoup plus précis pour des données du Mexique, comparativement au modèle logit. Nous reconnaissons toutefois avec Fay qu'il ne faudrait pas appliquer les modèles de Box-Cox avant d'avoir envisagé d'autres possibilités, comme la transformation des variables explicatives. Comme le souligne Fay, la méthode de Box-Cox sera utile dans les cas où elle pourra produire un modèle additif sur l'échelle transformée tandis que le modèle logit nécessiterait des termes d'interaction.

(ii) Réponse aux commentaires de E.A. Molina

Molina a raison de dire que l'on peut utiliser des mesures d'association pour examiner préalablement de nombreuses classifications combinées au coût le moins élevé possible. L'étude qu'il a réalisée en collaboration avec T.M.F. Smith, et dans laquelle il étend la théorie classique des mesures d'association aux données d'enquête recueillies à l'aide d'un plan avec échantillonnage en grappes et stratification, a contribué grandement à l'avancement des connaissances dans le domaine. Comme nous l'avons mentionné dans l'introduction, l'utilisateur est supposé connaître entièrement la matrice des covariances estimées des estimations par case. Toutefois, il arrive souvent que l'on ne dispose pas de ce genre d'information pour les analyses secondaires et, comme le souligne Molina, il peut même arriver que l'on ne connaisse pas les effets du plan par case. Rao et Scott (1987) ont montré que, dans un tel cas, un critère F utilisé dans GLIM pour tester une hypothèse emboîtée comme $\gamma = 0$ étant donné le modèle (2.12), est asymptotiquement valide

Ces commentaires portent en règle générale sur l'utilisation de fonctions de pseudo-vraisemblance. Le fait de ne pas tenir compte du plan de sondage peut se traduire par une augmentation ou une diminution de la variabilité prévue et ces fluctuations peuvent être intégrées dans un modèle de surdispersion ou de sous-dispersion à l'aide des fonctions de quasi-vraisemblance ou de leur version élargie. Voir à ce sujet Pocock et coll. (1981), Breslow (1984) et Williams (1982) notamment. À titre d'exemple, j'ai fait une nouvelle analyse des données du tableau 1. L'analyse faite par Kumar, Rao et Roberts est juste puisqu'elle comprend la matrice des covariances réelle. Supposons toutefois que nous ne connaissions pas cette matrice et que nous dispositions uniquement des effets du plan par case. À l'aide de GLIM, j'ai ajusté le modèle (2.12) avec erreur de Poisson et en ne tenant pas compte du plan d'échantillonnage, nous obtenons $X^2 = 5.68$, $G^2 = 5.67$. L'approximation de Rao et Scott (1987) pour le critère chi carré donne $X^2(\delta) = 5.68/2.25 = 2.52$. Pour le modèle d'indépendance, les valeurs non corrigées sont $X^2 = 18.22$, $G^2 = 18.22$, et la valeur corrigée est $X^2(\delta) = 18.22/1.65 = 11.04$. Que peut-on faire si on ne connaît pas les effets du plan? Une façon simple de procéder à l'aide d'une fonction de quasi-vraisemblance pour la surdispersion est d'estimer l'écart moyen pour le modèle le plus important, $D = 5.68/3 = 1.89$, et d'utiliser l'inverse de cette valeur comme poids (ou comme nouveau paramètre d'échelle). Nous obtenons ainsi $X^2 = 3.01$ pour le modèle (2.12) et $X^2 = 9.65$ pour le modèle d'indépendance. La façon juste de procéder, en l'occurrence, est d'utiliser l'excédent de la somme des carrés des écarts (différence entre les rapports de vraisemblance logarithmiques) pour tester $\gamma = 0$, puisque G^2 égalera le nombre de degrés de liberté pour le modèle le plus important. La valeur pertinente est 6.65, ce qui est tout juste significatif à un seuil de 1 %. Les deux analyses sont conformes à celle faite par les auteurs mais il pourrait en être autrement dans d'autres circonstances. Le modèle de quasi-vraisemblance exposé ici revient à supposer que la matrice des covariances réelle est un multiple de celle obtenue selon un échantillonnage multinomial; ce genre de modèle peut s'avérer inefficace dans plusieurs situations mais il a pour avantage de pouvoir être utilisé lorsque seule la variabilité inhérente des données est connue et que l'analyse est exécutée à l'aide d'un logiciel standard comme GLIM. Si les effets du plan sont connus, on peut proposer d'autres modèles qui en tiennent compte; d'ailleurs, on est à rédiger un article à ce sujet.

Néanmoins, on n'a pas encore réussi à trouver une forme d'analyse qui remplace parfaitement l'analyse fondée sur la matrice des covariances réelle. J'ai voulu ici mettre en évidence de nouvelles formes d'analyse à envisager lorsqu'on ne connaît pas entièrement la matrice des covariances. Les fonctions de quasi-vraisemblance sont un moyen tout indiqué de poursuivre la recherche, particulièrement en ce qui concerne les données d'enquête. L'article de Kumar, Rao et Roberts présente plusieurs avenues et, à ce titre, contribue notablement à l'avancement des connaissances dans le domaine.

SOURCES ADDITIONNELLES

- JØRGENSEN, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B* 127-162.
- MOLINA, E.A., et SMITH, T.M.F. (1986). The effect of sample design on the comparison of associations. *Biometrika* 73, 23-33.
- MOLINA, E.A., et SMITH, T.M.F. (1988). The effect of sampling on operative measures of association. *International Statistical Review* 56, 235-242.
- MORRIS, C.N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics* 10, 65-80.
- NELDER, J.A., et PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* 74, 221-232.
- POCOCK, S.J., COOK, D.G., et BERESFORD, S.A.A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Applied Statistics* 31, 286-295.
- WILLIAMS, D.A. (1982). Extra binomial variation in logistic-linear models. *Applied Statistics* 31, 144-148.

COMMENTAIRES

E.A. MOLINA¹

Les auteurs méritent des félicitations pour avoir brossé un tableau des méthodes qui ont été élaborées ces dernières années afin d'analyser les données qualitatives tirées d'enquêtes par sondage. Leur article devrait être très utile aux analystes d'enquêtes qui veulent prendre en considération l'effet des plans de sondage sur les aspects pratiques de l'analyse de données d'enquête. De façon plus particulière, il importe de souligner que les méthodes étudiées dans cet article se rapportent à deux types d'analyse: soit l'*analyse primaire*, pour laquelle le statisticien dispose de tous les renseignements pertinents, et l'*analyse secondaire*, pour laquelle le statisticien ne dispose pas de tous les renseignements voulus sur les unités de la population pour calculer dans son entier la matrice des covariances des estimateurs de l'échantillon.

Les méthodes analysées exigent la création d'un modèle structurel pour les données. Or, il est parfois difficile de construire un modèle structurel qui décrive convenablement des données qualitatives. Dans les grandes enquêtes, on doit souvent examiner préalablement de nombreuses classifications combinées au coût le moins élevé possible et pour cela, on aura tendance à utiliser des mesures d'association. Molina et Smith (1986, 1988) ont appliqué ces méthodes non paramétriques à des données d'enquête par sondage.

En ce qui concerne l'analyse primaire des données d'enquête, l'article met l'accent sur les moindres carrés pondérés et les tests de Wald. Les auteurs font un résumé des observations et la quasi-vraisemblance. Je crois que cette section de l'article devrait contenir une conclusion importante de l'analyse des auteurs, à savoir la nécessité de prendre en considération les contrastes d'enquête $K'p(X\beta) = \pi$ lorsqu'on utilise des méthodes de quasi-vraisemblance. Le lecteur n'est peut-être pas conscient de l'importance de bien choisir l'inverse de g dans l'équation (2.9). Les méthodes de quasi-vraisemblance sont maintenant d'usage courant et leur rapport avec les méthodes fondées sur les moindres carrés pondérés est indéniable. De fait, les fonctions de quasi-vraisemblance sont une solution de remplacement intéressante pour l'analyse de données d'enquête. Cependant, l'utilisation de ces fonctions n'est pas sans difficulté puisqu'il faut définir, en l'occurrence, la matrice des covariances comme une fonction de p , la fonction de variance. Les fonctions de quasi-vraisemblance sont largement déterminées par ces fonctions de variance (voir, par exemple, Morris 1982 et Jorgensen 1987). Si nous avons une matrice d'estimations au lieu d'une fonction, la méthode équivalait à utiliser une distribution normale.

L'article porte surtout sur des méthodes qui utilisent des fonctions de *pseudo-vraisemblance*. Comme l'analyse secondaire est celle que l'on retrouve le plus souvent en pratique, les méthodes exposées dans l'article sont susceptibles d'être largement utilisées par les analystes d'enquêtes. J'aimerais toutefois jeter un regard sur d'autres méthodes.

L'étude de l'incidence du plan de sondage sur les modèles de transformation de Guerrero et Johnson (1982) apporte beaucoup à la littérature existante. Or, Nelder et Pregibon (1987) proposent une famille de fonctions, dites *fonctions de quasi-vraisemblance élargies*, qui n'ont pas les principales faiblesses des modèles de transformation et qui peuvent être ajustées au moyen du logiciel GLIM. Si les effets du plan sont connus, on peut adapter ces méthodes aux données d'enquête en intégrant celles-ci aux fonctions de variance ou en les utilisant comme poids. Par ailleurs, les variables du plan peuvent servir à redresser les paramètres de dispersion dans les modèles. L'avantage dans les deux cas est que nous pouvons nous servir des critères utilisés dans les tests de validité de l'ajustement et des erreurs types calculées par GLIM selon ces modèles pour analyser les données, sans devoir effectuer de nouvelles corrections.

¹ E.A. Molina, Universidad Simon Bolívar, Caracas et University of Southampton, Royaume-Uni.

- C1 **Adaptabilité des méthodes multinomiales aux plans de sondage complexes:** La méthode MCP semble s'adapter plus facilement.
- C2 **Efficience:** Suivant un échantillonnage multinomial, la méthode MCP est, en général, asymptotiquement équivalente à la méthode PMV (il s'agit alors du MV ordinaire). On pourrait supposer que les MCP donneront toujours des estimations au moins aussi efficaces que celles du PMV suivant des plans de sondage complexes, bien que cette hypothèse présuppose une équivalence parfaite des deux méthodes d'estimation. Si la méthode MCP est plus efficiente, le gain d'efficience est-il habituellement négligeable (voir Scott et Holt 1982)? peut-on tirer de cela des résultats généraux?
- C3 **Degrés de liberté:** Les estimateurs par les MCP et les tests de Wald correspondants pourraient être instables si le nombre de degrés de liberté utilisé pour estimer V_p est faible.

SOURCES ADDITIONNELLES

FULLER, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexe. *Techniques d'enquête*, 10, 107-130.

SAS Institute Inc. (1985). *SAS/IML User's Guide, Version 5 Edition*. Cary NC: SAS Institute Inc.

SKINNER, C.J., HOLT, D., et SMITH, T.M.F., Eds. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

COMMENTAIRES

C.J. SKINNER¹

Dans leur article, Kumar, Rao et Roberts font une excellente analyse de l'application de la méthode des moindres carrés pondérés (MCP) et de la méthode du pseudo-maximum de vraisemblance (PMV) à des données qualitatives. Souhaitons que cet article, par sa présentation claire et ses exemples concrets, incitera les analystes d'enquêtes à considérer les plans de sondage complexes dans leurs analyses. Comme l'affirment les auteurs, les méthodes d'analyse statistiques qui tiennent compte des plans de sondage complexes ont fait l'objet de nombreuses recherches ces dernières années (voir, par exemple, Skinner, Holt et Smith 1989) et on commence même à les intégrer dans des logiciels standard (par ex.: SAS 1985, p. 61-67).

J'aimerais tout d'abord faire des commentaires sur des aspects particuliers de l'article. La section sur les variables polytomiques (section 5) est particulièrement utile, compte tenu du grand nombre d'enquêtes où l'on retrouve ce genre de données. Par définition, on s'attend souvent qu'une variable ordinaire soit en relation monotone avec d'autres variables; c'est pourquoi l'absence de monotonie entre les valeurs ajustées de $C_{1(i,k)}$ (ou $C_{2(i,k)}$) et la variable k dans le tableau 3 nous porte à croire que le résultat du test corrigé (à savoir qu'il n'y a aucun effet lié au niveau de scolarité) est plus plausible que celui du test non corrigé.

Le sujet traité à la section 4 (test d'équivalence de deux modèles de régression logistique) m'a aussi semblé d'une utilité concrète, bien qu'il serait possible théoriquement de définir ce test comme un test d'hypothèse emboîtée selon le modèle de Roberts, Rao et Kumar (1987). La section 3 montre très bien comment on peut appliquer la méthode PMV à des modèles paramétriques généraux pour données qualitatives. Il est toutefois agréable de constater que le modèle de régression logistique ne fournit pas vraiment un ajustement moins précis que celui qu'on obtient avec le modèle de transformation, plus complexe, ceci étant dû au fait que les paramètres du modèle de transformation sont plus difficiles à interpréter. Par exemple, le coefficient du niveau de scolarité dans le modèle logistique peut signifier que, pour chaque année de scolarité additionnelle chez les hommes de tous âges, la probabilité relative d'avoir une occupation augmentée de 16 % ($\exp(.1509) = 1.16$). Ce genre d'interprétation n'est habituellement pas possible pour le modèle de transformation lorsque $\lambda \neq 0$.

Sur un plan plus général, je voudrais connaître l'opinion des auteurs en ce qui a trait aux avantages relatifs des méthodes MCP et PMV. Dans leur article, ils présentent ces méthodes séparément mais celles-ci pourraient vraisemblablement s'appliquer à de très nombreux modèles pour données qualitatives provenant d'enquêtes complexes. De fait, les deux méthodes peuvent aussi s'appliquer à des modèles comprenant des variables continues (Skinner, Holt et Smith 1989, Chapitre 3); dans le cas des MCP, il faut simplement une statistique qui convienne à une fonction connue des paramètres ainsi qu'une estimation convergente de la matrice des covariances de cette statistique (Fuller 1984, Corollaire 2), tandis que dans le cas du PMV, nous avons vu dans Binder (1983) que ses applications sont vastes. À des fins d'analyse, j'énumère ci-dessous un certain nombre de critères par rapport auxquels les deux méthodes pourraient être comparées; les critères MI à M3 sont aussi valables pour un échantillonnage multinominal tandis que les critères C1 à C3 se rapportent uniquement à des plans de sondage complexes.

M1 Souplesses: La méthode MCP est peut-être plus souple que la méthode PMV pour des cas complexes comme ceux où il est question de zéros structurels.

M2 Calcul: La méthode MCP implique généralement un mode de calcul plus courant.

M3 Efficacités de case peu élevées: La méthode MCP est plus sensible aux faibles efficacités, particulièrement les efficacités nuls.

¹ C.J. Skinner, University of Southampton, Royaume Uni.

la transformation des variables x . Personnellement, je pencherais pour une analyse sur une échelle logistiqua avec, peut-être, des variables explicatives transformées, à moins que la transformation de Box-Cox ne présente un avantage marqué, comme l'existence d'un modèle additif sur l'échelle transformée lorsque le modèle logistique ne fournit pas un ajustement aussi précis en l'absence de termes d'interaction.

Je suis ravi de l'occasion qui m'est offerte de féliciter les auteurs pour un article aussi utile qu'instructif.

COMMENTAIRES

ROBERT E. FAY¹

L'article de Kumar, Rao et Roberts vient enrichir la littérature qui existe déjà sur l'analyse de données d'échantillons complexes. En analysant consécutivement quatre modèles pour données qualitatives, à savoir un modèle log-linéaire pour une classification combinée, un modèle de transformation de Box-Cox modifié pour les données binaires, une méthode d'inférence pour les paramètres d'un modèle de régression logistique, et un modèle de réponse polytomique, les auteurs proposent des réponses à des problèmes majeurs et montrent comment ces méthodes d'inférence peuvent être tout aussi bien appliquées à d'autres modèles pour données qualitatives tirées d'échantillons complexes. Les applications sont reliées entre elles par une théorie fondamentale, exposée en majeure partie dans Rao et Scott (1984), mais cet article a le mérite d'exposer plus en détail les implications de la théorie générale pour des modèles particuliers.

J'aimerais toutefois souligner un point que les auteurs semblent avoir négligé involontairement; en effet, on peut aussi envisager la répétition d'échantillons pour chacun des modèles analysés dans cet article. Du reste, cette méthode peut s'avérer parfois plus appropriée. Par surcroît, lorsque les premier, deuxième et quatrième modèles analysés s'appliquent à des données recoupées, on peut recourir à une version complète de la théorie de la répétition d'échantillons. Dans chaque cas, on se sert du test chi carré avec jackknife (Fay 1985) pour vérifier la validité de l'ajustement et comparer des modèles emboîtés et on calcule les erreurs types des paramètres par la méthode de la répétition d'échantillons.

La méthode de la répétition permet aussi de déterminer les erreurs types et les covariances des paramètres de modèles de régression logistique (voir section 4), ce qui permet dans certains cas de tester l'égalité de deux séries de paramètres de régression au moyen d'un test de Wald. En outre, le "jackknife" pourrait, semble-t-il, être utilisé avec le test du rapport des vraisemblances ou le test chi carré dans les cas où il y aurait des variables continues; cependant, il est indispensable de prouver cette hypothèse avant de recommander l'application de cette méthode. Si je fais valoir la méthode de la répétition comme solution de remplacement pour résoudre les problèmes exposés dans cet article, ce n'est pas pour donner à entendre qu'elle est supérieure du point de vue méthodologique aux méthodes de Rao et Scott (1984) mais simplement pour préciser qu'elle est un moyen de plus de résoudre les problèmes d'inférence signalés dans cet article et les autres problèmes du même genre. Par exemple, l'utilisation de plus en plus fréquente de la méthode de la répétition pour estimer la variance dans les enquêtes démographiques courantes du U.S. Bureau of the Census ouvre la voie à des analyses comme celles présentées dans cet article.

Je tiens également à souligner que les méthodes exposées dans l'article et les méthodes analogues découlant de la théorie de la répétition débordent le cadre de l'inférence fondée sur un plan de sondage complexe, qui est le sujet de l'article. On peut parler par exemple de l'imputation multiple ou des méthodes connexes, qui visent à tenir compte de la variabilité due aux données manquantes. Il est possible d'intégrer ce genre de variabilité à la définition que l'on a de la variance dans le cadre de l'inférence fondée sur un plan sans devoir modifier les méthodes exposées dans l'article. Les méthodes générales peuvent aussi s'appliquer à des cas d'inférence pour expériences complexes, où le plan pose des problèmes d'échantillonnage en grappes ou de stratification semblables à ceux que l'on retrouve dans les enquêtes à plan de sondage complexe.

Néanmoins, compte tenu des quatre modèles analysés, j'estime que l'on ne devrait pas utiliser le modèle de transformation de Box-Cox avant d'avoir considéré d'autres avenues, comme

¹ Robert E. Fay, U.S. Bureau of the Census, Washington, D.C. 20233.

- SCOTT, A.J., et HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- SCOTT, A.J., RAO, J.N.K., et THOMAS, D.R. (1989). Weighted least squares and quasi maximum likelihood estimation for categorical data under generalized linear models. *Linear Algebra and its Applications*, second special issue on Linear Algebra and Statistics, à paraître.
- SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Document de travail N° SSM-D 86-002, Statistique Canada.
- SINGH, A.C., et KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-257.
- SKINNER, C.J., HOLMES, D.J., et SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.
- STATISTIQUE CANADA (1977). *Méthodologie de l'enquête sur la population active du Canada*, 1976. N° 71-526 au catalogue hors série, Statistique Canada.
- THOMAS, D.R., et RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

BIBLIOGRAPHIE

- AMEMIYA, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.
- BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BOX, G.E.P., et COX, D.R. (1964). An analysis of transformations (avec discussion). *Journal of the Royal Statistical Society, Sér. B*, 26, 211-252.
- BOX, G.E.P., et COX, D.R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209-210.
- FAY, R.E. (1985). A jack-knifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā, Sér. C*, 37, 117-132.
- FULLER, W.A. (1986). Estimators of the factor model for survey data. Dans *Advances in the Statistical Sciences*, Vol. I (Eds. MacNeill, I.B. and Umphey, G.J.). Dordrecht, Holland: Reidel Publishing Co., 265-284.
- GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society, Sér. B*, 46, 270-272.
- GUERRERO, V.M., et JOHNSON, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, 69, 309-314.
- HABERMAN, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.
- HIDIROGLOU, M.A., et RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys, Parties I et II. *Journal of Official Statistics*, 3, 117-132 et 133-140.
- HINKLEY, D.V., et RUNGER, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302-309.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples (avec discussion). *Journal of the Royal Statistical Society, Sér. B*, 36, 1-37.
- KOCH, G.G., FREEMAN, D.H. Jr., et FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KUMAR, S., et RAO, J.N.K. (1985). Fitting Box-Cox transformation models to labour force survey data. Rapport interne, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa.
- McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Sér. B*, 42, 109-142.
- NATHAN, G., et HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the American Statistical Association*, 76, 681-689.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., et SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- ROBERTS, G. (1985). *Contributions to Chi-Squared Tests with Survey Data*. Thèse de doctorat non publiée, Carleton University, Department of Mathematics and Statistics, Ottawa.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SCOTT, A.J. (1986). Logistic regression analysis with survey data. *Proceedings of the Section on Survey*

Tableau 4

Critères utilisés dans les tests de validité de l'ajustement
et les tests d'hypothèses emboîtées

Validité de l'ajustement (âge et niveau de scolarité)		Hypothèse emboîtée (âge seulement)	
X^2	37.7		7.1
X^2_c	21.6		3.8
X^2_{δ}	18.5*		3.7*
δ	1.75		1.9
a^2	0.83		0.1

* On a corrigé le critère de Satterthwaite de manière à pouvoir le comparer à la même valeur de chi carré χ^2 que X^2_c .

où $C_j(ik)$ est la probabilité cumulative j pour le groupe d'âge i et le groupe de niveau de scolarité k . En outre, $a_i = A_i - \bar{A}$, où A_i est le milieu de la tranche d'âge i et e_k est l'effet du groupe de niveau de scolarité k ($\sum e_k = 0$), si on ne tient pas compte de l'ordre des catégories. Le tableau 3 contient les proportions cumulatives estimées résultant de l'enquête tandis que le tableau 4 contient les critères X^2 , X^2_c et X^2_{δ} utilisés pour tester la validité de l'ajustement du modèle (5.15) de même que l'hypothèse emboîtée de l'absence d'effet pour le niveau de scolarité, $e_k = 0$ pour $k = 1, 2$.

Premièrement, en ce qui concerne la validité de l'ajustement du modèle (5.15), si nous ne tenons pas compte du plan de sondage et que nous comparons la valeur de X^2 à $\chi^2_{0.05}(13) = 22.4$, qui est la limite supérieure (à 5 %) de la distribution chi carré χ^2 avec $1I - 5 = 13$ d.l., nous devons rejeter le modèle. Par contre, si nous comparons la valeur de X^2 ou la valeur de X^2_{δ} une fois corrigée, à $\chi^2_{0.05}(13)$, nous voyons qu'elle n'est pas significative à un seuil de 5 %; par conséquent, le modèle fournit dans ce cas un ajustement approprié pour les données.

Pour ce qui a trait à l'hypothèse emboîtée, lorsque nous comparons la valeur de X^2 ou la valeur de X^2_{δ} , une fois corrigée, à $\chi^2_{0.05}(2) = 5.99$, nous voyons qu'elle n'est pas significative à un seuil de 5 %, ce qui confirme l'hypothèse emboîtée qu'il n'y a pas d'effet attribuable à l'éducation.

6. LOGICIELS

Pour appliquer les méthodes exposées dans cet article, il a fallu exécuter deux séries de calculs: premièrement, calcul d'un vecteur de proportions et de la matrice des covariances correspondante et deuxièmement, calcul des estimations de modèle, des critères utilisés dans un test et de leurs valeurs corrigées.

Des enquêtes comme l'Enquête Santé Canada et l'enquête sur la population active, où nous avons puise nos exemples, sont caractérisées par des plans de sondage complexes et de vastes bases de données. À cause de cela, nous avons dû calculer les matrices des covariances à l'aide d'un gros ordinateur, en nous servant de programmes SAS et Fortran conçus spécialement à cette fin.

Pour ce qui a trait aux opérations de lissage et aux tests de validité de l'ajustement et de sous-hypothèses, les calculs pertinents ont été effectués soit sur un gros ordinateur avec un programme SAS (et, notamment, le logiciel MATRIX) ou sur un micro-ordinateur avec le logiciel GAUSS. Ces programmes sont à la disposition des analystes de Statistique Canada.

Dans l'équation ci-dessus, $a(2|1)^2$ est déterminé en remplaçant θ par $(\hat{\theta}; 0')$ dans la définition suivante:

(5.13)
$$a(2|1)^2 = \left\{ \sum_n^{i=1} \delta_i(2|1)^2 - n\hat{u}_0.(2|1)^2 \right\} / n\hat{u}_0.(2|1)^2,$$

où

(5.14)
$$\sum_n^{i=1} \delta_i(2|1)^2 = \text{tr}D(2|1)^2.$$

On pourra définir les versions corrigées du test de validité de l'ajustement X^2 comme des cas particuliers de (5.9) et de (5.12) en considérant le modèle comme emboîté dans un modèle saturé (c.-à-d. un modèle où le paramètre inconnu θ est de dimension lJ).

Exemple

Les méthodes exposées ci-dessus ont été appliquées à des données de l'Enquête Santé Canada de 1978-1979. Cette enquête est décrite brièvement dans la section 2.

Les données analysées portaient sur une population de femmes âgées de 20 à 64 ans, répartie selon trois critères: fréquence d'auto-examen des seins (3 catégories: mensuellement, trimestriellement, moins fréquemment ou jamais), niveau de scolarité (3 catégories: études secondaires ou moins, études post-secondaires non complétées, études post-secondaires complétées) et âge (3 catégories: 20-24, 25-44, 45-64).

La fréquence d'auto-examen des seins était définie comme la variable de réponse tandis que le niveau de scolarité et l'âge représentaient les variables explicatives, de sorte que le nombre de choix de réponses, $J + 1$, s'élevait à 3 et le nombre de domaine I , à 9. La variable de réponse et les variables explicatives sont toutes trois ordinales.

Nous avons alors considéré un modèle semblable à celui décrit dans l'équation (5.4) pour les probabilités cumulatives:

(5.15)
$$\log\{C_j(ik)/(1 - C_j(ik))\} = v_j + \beta a_i + e_k \quad (j = 1, 2, 3; i = 1, 2, 3; k = 1, 2, 3)$$

Tableau 3

Valeurs estimées des probabilités cumulatives, selon les données de l'enquête

Age	Niveau de scolarité	$C_{1(ik)}$	$C_{2(ik)}$
$i = 1, k = 1$			
20-24			
\leq Etudes secondaires		.25	.49
$<$ Etudes post-secondaires non complétées		.25	.41
\geq Etudes post-secondaires complétées		.23	.47
$i = 1, k = 2$			
25-44			
\leq Etudes secondaires		.25	.50
$<$ Etudes post-secondaires non complétées		.27	.44
\geq Etudes post-secondaires complétées		.26	.44
$i = 1, k = 3$			
45-64			
\leq Etudes secondaires		.28	.51
$<$ Etudes post-secondaires non complétées		.24	.62
\geq Etudes post-secondaires complétées		.29	.56
$i = 3, k = 1$			

Correction des tests ordinaires

Pour des raisons de simplicité, nous n'utiliserons que le critère chi carré de Pearson pour tester la validité de l'ajustement du modèle (5.1). Ce critère est défini

$$X^2 = n \sum_{i=1}^I \sum_{j=1}^J w_i (f_j^{(i)} - F_j)^2 / F_j. \quad (5.7)$$

Suivant un échantillonnage multinomial indépendant dans chacun des domaines, nous savons que X^2 est distribué asymptotiquement suivant la loi de chi carré χ^2 avec $II - r$ d.l. Pour ce qui a trait au test de l'hypothèse emboîtée $\theta_2 = 0$, étant donné le modèle (5.1) définissons θ_1 comme le pseudo-e.m.v. de θ_1 et F comme le vecteur correspondant des proportions de réponse ajustées, où $\theta' = (\theta_1', \theta_2') = (\theta_1', \theta_2')$, θ_1 est de dimension $q \times 1$ et θ_2 est de dimension $n \times 1$ ($q + n = r$). Le test chi carré de Pearson pour l'hypothèse emboîtée est alors défini par l'expression

$$X^2(2|1) = n \sum_{i=1}^I w_i \sum_{j=1}^J (F_{ij} - \hat{F}_{ij})^2 / \hat{F}_{ij}. \quad (5.8)$$

La variable $X^2(2|1)$ est distribuée asymptotiquement suivant la loi de chi carré χ^2 avec n d.l. selon un échantillonnage multinomial indépendant dans chacun des domaines. Toutefois dans le cas d'un plan de sondage général, $X^2(2|1)$ sont toutes deux distribuées asymptotiquement comme des sommes pondérées de variables chi carré χ^2 indépendantes avec, dans chaque cas, 1 d.l., les poids pouvant, en l'occurrence, être considérés comme des "effets du plan généralisés" de transformations linéaires particulières de F (Roberts 1985).

On obtient une correction du premier degré de $X^2(2|1)$ en considérant

$$X^2_2(2|1) = X^2(2|1) / \delta_1(2|1) \text{ comme une variable chi carré } \chi^2 \text{ avec } n \text{ d.l.} \quad (5.9)$$

Dans l'équation ci-dessus, $\delta_1(2|1)$ est déterminé en remplaçant θ' par $(\theta_1', 0')$ et V_p par V_p dans la définition suivante de $\delta_1(2|1)$:

$$u\delta_1(2|1) = \sum_{i=1}^I \delta_i(2|1) = \text{tr } D(2|1). \quad (5.10)$$

Dans l'équation (5.10), tr désigne l'opérateur trace et $D(2|1)$ est une matrice d'"effets du plan généralisés" définie par l'expression

$$D(2|1) = (H_2' \nabla H_2)^{-1} (H_2' \nabla V_p \nabla' H_2), \quad (5.11)$$

où V_p est la matrice des covariances de P , $\nabla = (D(W) \otimes I) \tilde{Q}^{-1}$, \tilde{Q} étant la matrice quasi-diagonale, avec $\tilde{Q}_i = \text{diag}(F_i) - F_i F_i'$, $i = 1, \dots, I$, $F_i = F_i(\theta)$, et $H_2 = [I - M_1 (M_1' \nabla M_1)^{-1} M_1' \nabla] M_2$, où $M_1 = (\partial F / \partial \theta_1)'$ et $M_2 = (\partial F / \partial \theta_2)'$.

On obtient une correction du second degré (plus précise) de $X^2(2|1)$, fondée sur l'approximation de Satterthwaite, en considérant

$$X^2_3(2|1) = X^2_2(2|1) / [1 + \hat{a}(2|1)^2] \text{ comme une variable } \chi^2 \text{ avec } n/[1 + \hat{a}(2|1)^2] \text{ d.l.} \quad (5.12)$$

(5.2)
$$F_{ij}(\theta) = \exp(x'_i \beta_j) \left/ \sum_{j=1}^J \exp(x'_i \beta_k) \right., \quad i = 1, \dots, I; \quad j = 1, \dots, J + 1.$$

et $\sum \beta_k = 0$. A cause de la contrainte s'appliquant aux β_k , nous pouvons réécrire l'équation (5.2) comme suit:

$$F_{ij}(\theta) = \exp(x'_i \beta_j) \left/ \left[\sum_{j=1}^k \exp(x'_i \beta_k) + \prod_j \exp(-x'_i \beta_k) \right] \right.,$$

(5.3)
$$i = 1, \dots, I; \quad j = 1, \dots, J.$$

Il convient de souligner que l'équation (5.3) se ramène à l'équation du modèle de régression logistique ordinaire lorsque la variable de réponse est binaire. En ce qui concerne les variables de réponse ordinales, McCullagh (1980) propose un modèle simple qui a la propriété d'être invariant lorsque des catégories de réponse sont groupées:

(5.4)
$$\log \{C_{j(i)} / (1 - C_{j(i)})\} = v_j - x'_i \beta, \quad j = 1, \dots, J; \quad i = 1, \dots, I$$

où $C_{j(i)} = \sum_{k=1}^j P^{k(i)}$ désigne la probabilité cumulative j dans le domaine i et $\theta' = (v_1, \dots, v_J, \beta')$. Si nous voulons exprimer l'équation (5.4) sous la forme définie en (5.1), notons que $P_i = L^{-1}C_i$, où $P_i = (P_{1(i)}, \dots, P_{J(i)})'$, $C_i = (C_{1(i)}, \dots, C_{J(i)})'$ et L^{-1} est une matrice régulière $J \times J$ constituée de la façon suivante: valeur 1 pour les éléments diagonaux, valeur -1 pour les éléments $(i + 1, i)$ ($1 < J$) et valeur 0 pour les autres éléments.

Pseudo-EMV

Comme dans les deux sections précédentes, nous utilisons ici des pseudo-e.m.v., θ , que nous avons tirés des équations de vraisemblance de θ selon un modèle multinomial produit en remplaçant les proportions de réponse simples n_{ij}/n_i par les estimations d'enquête correspondantes $F_{j(i)}$, et en remplaçant n_i/n par l'estimation d'enquête correspondante W_i de la proportion de domaine W_i . Ici, n_{ij} représente le nombre d'unités ayant donné la réponse j dans un échantillon de taille n_i du domaine i et $n = \sum n_i$. Les proportions de réponse ajustées sont alors définies par l'équation $F = F(\theta) = (F'_1, \dots, F'_J)'$, où $F'_i = (F'_{i1}, \dots, F'_{ij})'$ et $F'_{ij} = F_{ij}(\theta)$. Soit V_p la matrice des covariances estimée des estimations d'enquête $F = (P_{1(1)}, \dots, P_{J(1)})', \dots, P_{1(J)}, \dots, P_{J(J)})'$, et $\hat{M} = (\partial F / \partial \theta)'$, la matrice $IJ \times r$ des dérivées partielles $\partial F_{ij} / \partial \theta_k$ évaluées à θ . De plus, soit $\hat{Q}_i = \text{diag} (F'_i) - F'_i F'_i$ et $\hat{Q} = \text{diag} (\hat{Q}_i, i = 1, \dots, I)$. Les expressions des dérivées partielles $\partial F_{ij} / \partial \theta_k$ pour les modèles (5.3) et (5.4) sont reproduites dans Roberts (1985). Nous définissons alors comme suit la matrice des covariances asymptotique estimée de θ , compte tenu du plan de sondage (voir Roberts 1985):

(5.5)
$$\text{est cov}(\theta) = (M' \hat{\Delta} M)^{-1} (M' \hat{\Delta} V_p \hat{\Delta} M) (M' \hat{\Delta} M)^{-1},$$

où $\hat{\Delta} = (D(W) \otimes I) \hat{Q}^{-1}$ et $D(W) = \text{diag} (W_i, i = 1, \dots, I)$. Lorsqu'il s'agit plus précisément d'un échantillonnage multinomial produit, $V_p = \hat{\Delta}^{-1}/n$ et l'équation (5.5) se réduit à $(M' \hat{\Delta} M)^{-1}/n$.

Le vecteur des résiduels, $R = P - F$, présente aussi un certain intérêt car il peut servir à mettre en évidence les faiblesses du modèle. La matrice des covariances asymptotique estimée de R est définie

(5.6)
$$\text{est cov}(R) = G' P G, \quad \text{ou } G = I - M(M' \hat{\Delta} M)^{-1} M' \hat{\Delta}.$$

Exemple

La méthode exposée ci-dessus a été appliquée à des données de l'enquête sur la population active d'octobre 1980 et d'octobre 1981 afin d'analyser les changements structurels survenus d'une année à l'autre.

Le modèle de régression logistique avec effets linéaire et quadratique pour l'âge et effet linéaire pour le niveau de scolarité fournissait un ajustement approprié pour les données des deux périodes avec les valeurs estimées de β_j suivantes:

$$\beta_1: \{-3.08, 0.211, -0.00218, 0.1505\}$$
$$\beta_2: \{-3.05, 0.179, -0.00169, 0.1707\},$$

où $\log\{F_{ijk}/(1 - F_{ijk})\} = \beta_{i0} + \beta_{i1}A_j + \beta_{i2}A_j^2 + \beta_{i3}E_k, j = 1, \dots, 10; k = 1, \dots, 6$ et F_{ijk} est le taux d'emploi ajusté dans la case (j, k) pour la période t . Une case a été exclue du processus d'ajustement parce que la taille d'échantillon de domaine n_{jt} est nulle pour la période courante.

Pour ce qui a trait au test de l'hypothèse $\beta_1 = \beta_2$, nous avons obtenu les valeurs suivantes pour X^2, G^2, X^2_c, G^2_c et X^2_{\S}, G^2_{\S} étant donné les modèles de régression logistique:

$$\begin{array}{llll} X^2 = 42.1 & X^2_c = 24.6 & X^2_{\S} = 24.4 \\ G^2 = 42.2 & G^2_c = 24.6 & G^2_{\S} = 24.4 \end{array}$$

Par ailleurs, $s/(1 + \hat{\sigma}^2) = 4/(1.0089) = 3.965 \doteq 4$. Si nous comparons la valeur de X^2_{\S} ou de G^2_{\S} à $\chi^2_{0.05}(4) = 9.49$, qui est la limite supérieure (à 5 %) de la distribution chi carré avec 4 d.l., nous rejetons l'hypothèse $\chi^2 \beta_1 = \beta_2$ à un seuil de 5 % et en concluons qu'il y a eu des changements structurels significatifs entre octobre 1980 et octobre 1981. Par conséquent, on ne peut utiliser les données des deux périodes pour établir des estimations lissées des taux de chômage, $1 - F_{jk}$, pour la période courante.

5. MODELES DE REPONSE POLYTOMIQUE

Les ouvrages économétriques proposent une quantité de modèles pour le cas où la variable de réponse est polytomique. Cette pluralité reflète en partie les diverses échelles de mesure qu'il peut y avoir pour les variables de réponse polytomiques, contrairement aux variables de réponse binaires. De façon générale, il y a les variables de réponse qualitatives, où l'ordre des catégories de réponse importe peu, et les variables de réponse ordinales, où il existe un ordre naturel pour les catégories de réponse.

Supposons que la population d'intérêt est répartie en I cases (ou domaines) formées selon un ou plusieurs critères de classification. Soit $P_{f(i)}$ la proportion de population dans la case i pour la réponse f ($f = 1, \dots, J + 1$), de sorte que $\sum_{f=1}^{J+1} P_{f(i)} = 1$ ($i = 1, \dots, I$). Nous pouvons alors définir un modèle de réponse polytomique général pour les proportions $P_{fj}(i)$:

$$P_{fj}(i) = F_{fj}(\theta), \quad i = 1, \dots, I; \quad j = 1, \dots, J, \tag{5.1}$$

où θ est un r -vecteur de paramètres inconnus ($r \leq IJ$) et $F_{fj}(\theta)$ est une fonction de forme connue. En ce qui concerne les variables de réponse qualitatives, Haberman (1982) et d'autres proposent le modèle suivant: les "logits polynomiaux" $\log P_{fj}(i) - \sum_{f'=1}^{j+1} \log P_{f'}(i) (J + 1)^{-1}$ sont supposés être des fonctions linéaires inconnues de x_i , qui est le s -vecteur de constantes connues dérivé des niveaux de facteurs, c'est-à-dire,

Correction des tests ordinaires

Les versions standard du test chi carré et du test du rapport des vraisemblances pour l'hypothèse emboîtée $\beta_1 = \beta_2$, étant donné le modèle (4.1), sont définies respectivement par les équations

(4.8)
$$X^2 = X_1^2 + X_2^2$$

et

(4.9)
$$G^2 = G_1^2 + G_2^2,$$

où

(4.10)
$$X_t^2 = n_t \sum_I (F_{it} - \hat{F}_i)^2 W_{it} / \{ \hat{F}_i (1 - \hat{F}_i) \}, \quad t = 1, 2$$

et

(4.11)
$$G_t^2 = 2n_t \sum_I W_{it} \left[F_{it} \log(F_{it} / \hat{F}_i) + (1 - F_{it}) \log \{ (1 - F_{it}) / (1 - \hat{F}_i) \} \right], \quad t = 1, 2.$$

On obtient une correction du premier degré de X^2 (ou G^2) en considérant $X_c^2 = X^2 / \delta$, ou $G_c^2 = G^2 / \delta$, comme une variable chi carré avec s degrés de liberté, où χ^2

(4.12)
$$s\delta = n_1 \sum_I V_{11R}(ii) W_{1i} / \{ \hat{F}_i (1 - \hat{F}_i) \} + n_2 \sum_I V_{22R}(ii) W_{2i} / \{ \hat{F}_i (1 - \hat{F}_i) \}$$

et $V_{nr}(ij)$ est l'élément (ij) de V_{nr} . On obtient une correction du second degré (plus précise) de X_c^2 (ou de G_c^2), fondée sur l'approximation de Satterthwaite, en considérant

(4.13)
$$X_s^2 = \frac{X_c^2}{G_c^2} = \frac{1 + a^2}{G_c^2} \quad \text{ou} \quad G_s^2 = \frac{1 + a^2}{G_c^2} \quad \text{comme une } \chi^2 \text{ avec } s/(1 + a^2) \text{ d.l.}$$

Dans les équations ci-dessus, $a^2 = (\sum_{k=1}^s \delta_k^2 - s\delta^2) / s\delta^2$ et peut être calculé à l'aide de l'équation (4.12) et de la formule suivante $\sum \delta_k^2$:

(4.14)
$$\begin{aligned} \sum_s \delta_k^2 &= n_1^2 \sum_I \sum_{j=1}^I \frac{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)}{V_{11R}(ij) W_{1i} W_{2j}} \\ &+ n_2^2 \sum_I \sum_{j=1}^I \frac{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)}{V_{22R}(ij) W_{2i} W_{2j}} \\ &+ 2n_1 n_2 \sum_I \sum_{j=1}^I \frac{\hat{F}_i \hat{F}_j (1 - \hat{F}_i) (1 - \hat{F}_j)}{V_{12R}(ij) W_{1i} W_{2j}} \end{aligned}$$

où $V_{12R}(ij)$ est l'élément (ij) de V_{12R} .

Suivant l'hypothèse que $\beta_1 = \beta_2 (= \beta)$, les pseudo-estimations les plus vraisemblables, $\hat{\beta}$, sont calculées par itération au moyen des pseudo-équations de vraisemblance:

$$X'D(W_c)\hat{F} = (n_1/n)X'D(W_1)\hat{F}_1 + (n_2/n)X'D(W_2)\hat{F}_2, \tag{4.3}$$

où $D(W_c) = (n_1/n)D(W_1) + (n_2/n)D(W_2)$, $\hat{F} = F(\hat{\beta})$ est le vecteur des proportions de réponse ajustées ou des estimations lissées des proportions par case pour la période courante, et $n_1 + n_2 = n$.

Soit V_p la matrice des covariances estimée de $(\hat{F}_1, \hat{F}_2)'$ de la forme

$$V_p = \begin{bmatrix} V_{11p} & V_{12p} \\ V_{21p} & V_{22p} \end{bmatrix}.$$

Alors, la matrice des covariances estimée des estimations lissées \hat{F} est définie par l'expression

$$\text{est cov}(\hat{F}) = BV_pB', \tag{4.4}$$

où

$$B = D(W_c)^{-1}\hat{\Delta}X(X'\hat{\Delta}X)^{-1}X'[(n_1/n)D(W_1), (n_2/n)D(W_2)] \tag{4.5}$$

et

$$\hat{\Delta} = \text{diag}(W_c\hat{F}_i(1 - \hat{F}_i)), i = 1, \dots, I.$$

Si les résiduels sont définis par l'expression $R_i = \hat{F}_i - F_i$, la matrice des covariances estimée de $(R_1', R_2)'$ est

$$\hat{V}_R = \begin{bmatrix} \hat{V}_{11R} & \hat{V}_{12R} \\ \hat{V}_{21R} & \hat{V}_{22R} \end{bmatrix} = AV_pA'. \tag{4.6}$$

Dans l'équation ci-dessus,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

avec

$$A_{11} = D(W_c)^{-1}\hat{\Delta}X \left[(X'\hat{\Delta}'X)^{-1}X'D(W_1) - \frac{n}{n'}(X'\hat{\Delta}X)^{-1}X'D(W_1) \right],$$

et

$$A_{12} = -D(W_c)^{-1}\hat{\Delta}X(X'\hat{\Delta}X)^{-1}X' \left\{ D(W) - \frac{n}{n'}D(W_1) \right\}, \quad i = 1, 2,$$

ou

$$\hat{\Delta}_i = \text{diag}(W_i\hat{F}_i(1 - \hat{F}_i)), i = 1, \dots, I.$$

conditions, on peut déterminer la matrice des covariances estimée de β par l'équation (3.5) en remplaçant $\partial F/\partial \theta$ par $\partial F/\partial \beta$ dans l'expression pour B (équation (3.3)). Dans notre exemple, cela voudrait dire que nous pouvons opter pour $\lambda = 0$ et utiliser les valeurs estimées de β et les erreurs types correspondantes (ou la matrice des covariances estimée) calculées selon le modèle de régression logistique (voir tableau 2).

4. TEST D'ÉQUIVALENCE DE MODÈLES DE RÉGRESSION LOGISTIQUE

On peut découvrir les changements structurels survenus entre deux périodes par des tests d'égalité des paramètres des modèles correspondants. Ce genre de tests a été traité en profondeur dans les ouvrages d'économétrie en ce qui concerne les modèles de régression linéaire courants (voir par exemple Amemiya 1985, sec. 1.5.3).

Dans cette section, nous appliquons des versions corrigées du test chi carré et du test du rapport des vraisemblances pour tester l'égalité des paramètres de deux modèles de régression logistique qui se rapportent à deux périodes déterminées. Si l'hypothèse de l'égalité est valide, il est alors possible d'obtenir des estimations "lissées" (c.-à-d. ajustées) des proportions par case pour la période courante en combinant les données relatives aux deux périodes. Ces estimations sont plus efficaces que les estimations lissées fondées uniquement sur les données de la période courante. Nous allons appliquer ces méthodes à des données de l'enquête sur la population active d'octobre 1980 et d'octobre 1981 afin d'analyser les changements structurels d'une année à l'autre. Rappelons que nous avons déjà utilisé les données d'octobre 1980 dans la section 3 afin d'illustrer l'ajustement des modèles de transformation de Box-Cox et que nous avons constaté qu'un modèle de régression logistique avec effets linéaire et quadratique pour l'âge et effet linéaire pour le niveau de scolarité fournissait un ajustement approprié pour les données.

Soit P_{it} la proportion de réponse de la population dans la case i pour la période t ($t = 1, 2$). Nous définissons alors comme suit un modèle de régression logistique pour les proportions $P_{it} = F_i(\beta_t)$:

(4.1) $\log\{F_{it}/(1 - F_{it})\} = x_i'\beta_t, \quad i = 1, \dots, I; t = 1, 2$

où x_i est un s-vecteur de constantes connues dérivé des niveaux de facteur, comme en (3.1), et β_t est un s-vecteur de paramètres inconnus pour la période t . Nous voulons tester l'hypothèse composée $\beta_1 = \beta_2$ ($= \beta$) pour analyser les changements structurels survenus entre les deux périodes. Si l'hypothèse est acceptée, $F_i(\beta)$ où β est le pseudo-e.m.v. du paramètre commun β , représentera des estimations "lissées" des proportions P_{it} pour la période courante ($t = 2$).

Pseudo - EMV

Soit P_{it} et P_{2i} ($i = 1, \dots, I$) les estimations d'enquête fondées sur des échantillons de taille n_1 et n_2 respectivement. Comme dans la section 3, les "pseudo-estimateurs du maximum de vraisemblance", $\hat{\beta}_t$, sont tirés des équations de vraisemblance de β_t selon un modèle binomial pro-duit; nous obtenons ces pseudo-estimateurs en remplaçant les proportions de réponse simples r_{it}/n_{it} par les estimations d'enquête correspondantes P_{it} de P_{it} et en remplaçant n_{it}/n_t par les estimations d'enquête correspondantes W_{it} des proportions de domaine W_{it} , ce qui donne

(4.2) $X'D(W_t)F_t = X'D(W_t)P_t, \quad t = 1, 2$

où $F_t = F(\beta_t)$ est le vecteur des proportions de réponse ajustées pour la période t , $D(W_t) = \text{diag}(W_{1t}, \dots, W_{It})$, et $X' = (x_1', \dots, x_I')$. Les estimations $\hat{\beta}_t$ sont calculées de façon itérative à l'aide d'une quasi-méthode de Newton.

Tableau 2
Pseudo-estimations les plus vraisemblables des paramètres (β' , λ), erreurs types correspondantes et critères utilisés dans un test selon le modèle de transformation et le modèle de régression logistique correspondant ($\lambda = 0$)

	Modèle de transformation		Modèle de régression logistique	
	valetur estimée	e.t.	valetur estimée	e.t.
β_0	- 3.28	0.975	- 3.10	0.247
β_1	0.219	0.0468	0.211	0.013
β_2	- 0.00227	0.00049	- 0.00218	0.00017
β_3	0.1579	0.0385	0.1509	0.0115
λ	0.0160	.085	—	—
Critères utilisés dans un test				
	valetur	d.l.	valetur	d.l.
X^2	99.6	55	99.8	56
G_2	102.6	56	102.5	56
X^2_5	40.7	39.2	23.4	24.2
G^2_5	42.0	39.2	23.9	24.2
$X^2_5(0.05)$	54.6	55	47.7	56
$G^2_5(0.05)$	56.4	55	48.9	56

Le tableau 2 contient les pseudo-estimations les plus vraisemblables e.m.v. de $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \lambda)'$ et les erreurs types correspondantes de même que les critères X^2 , G^2 , X^2_5 et G^2_5 utilisés pour tester la validité de l'ajustement du modèle (3.12). À des fins de comparaison, nous avons aussi inclus dans ce tableau les valeurs correspondantes pour un modèle de régression logistique ($\lambda = 0$).

De toute évidence, la valeur de X^2 (or G^2) est essentiellement la même pour les deux modèles. Dans notre exemple, le modèle de transformation ne fournit donc pas un meilleur ajustement que le modèle de régression logistique. C'est aussi la conclusion que nous pouvons déduire de la valeur de λ ($= 0.016$) qui n'est pas réellement différente de $\lambda = 0$ lorsqu'on tient compte des mêmes pour les deux modèles mais les erreurs types des β_i sont beaucoup plus élevées selon le modèle de Box-Cox que selon le modèle de régression logistique parce que λ a une erreur type élevée et que les valeurs β_i dépendent de λ .

Si nous ne tenons pas compte du plan de sondage et que nous comparons la valeur de X^2 (ou de G^2) à $\chi^2_{0.05}(55) = 73.3$, qui est la limite supérieure (à 5 %) de la distribution chi carré χ^2 avec $I - s - 1 = 55$ d.l., nous rejèterons le modèle (3.12). Par contre, si nous corrigeons la valeur de X^2_5 (ou de G^2_5) de manière à la comparer à $\chi^2_{0.05}(55)$, (la valeur redressée étant désignée par $X^2_5(0.05)$ (ou $G^2_5(0.05)$) dans le tableau 2), nous voyons qu'elle n'est pas significative à un seuil de 5 %, ce qui indique que le modèle fournit un ajustement précis pour les données P_{ijk} . Box et Cox (1982) et Hinkley et Runger (1984) ont soutenu que l'inférence statistique portant sur β devrait se faire à l'échelle déterminée par l'estimation λ tenue pour fixe. Dans ces

(3.9)
$$(I - s - 1)\delta_k = \sum_I V_{III,R} W_I / \{F_I(1 - F_I)\}$$

et $V_{III,R}$ est la variance estimée du résiduel R_I .
On obtient une correction du second degré (plus préciser) de X^2 (ou de G^2), fondée sur l'approximation de Satterthwaite de $\sum \delta_k W_k$, en considérant

(3.10)
$$X^2_s = \frac{X^2_c}{1 + a^2} \text{ ou } G^2_s = \frac{G^2_c}{1 + a^2} \text{ comme des } \chi^2 \text{ avec } (I - s - 1)/(1 + a^2) \text{ d.l.}$$

Dans les expressions ci-dessus, $a^2 = \sum (\delta_k - \delta_c)^2 / \{(I - s - 1)\delta_c^2\}$ est le carré du coefficient de variation des δ_i que l'on peut calculer, sans devoir évaluer chacun des poids δ_i , à l'aide de l'équation (3.9) et de

(3.11)
$$\sum \delta_k^2 = \sum_I \sum_{l=1}^{I-1} V_{III,R}^2 (nW_l)(nW_l) / \{f_l f_l (1 - f_l)(1 - f_l)\},$$

ou $V_{III,R}$ est l'élément (i,l) de V_R définie en (3.6).
Étant donné le modèle (3.2), il est aussi possible de tester des hypothèses emboîtées en apportant les corrections voulues aux tests ordinaires; cependant, pour des raisons de simplicité, nous n'aborderons pas ici cette question (voir Roberts 1985 et Kumar et Rao, 1985 pour plus de détails). Il sera plus simple d'utiliser des tests de Wald fondés sur les estimations $\hat{\beta}$ la matrice des covariances asymptotique estimée correspondante.

Exemple

La méthode que nous venons d'exposer a été appliquée à des données de l'enquête sur la population active du Canada d'octobre 1980. Le plan de l'enquête sur la population active prévoit un échantillonnage en grappes à plusieurs degrés, notamment à deux degrés dans les régions urbaines autoremplissantes et à trois ou quatre degrés dans les régions non autoremplissantes de chaque province. On trouvera une description détaillée du plan d'échantillonnage et des méthodes d'estimation de l'enquête sur la population active dans Statistique Canada (1977). Dans notre exemple, l'échantillon de l'enquête sur la population active comprend des hommes âgés de 15 à 64 ans qui font partie de la population active et ne sont pas des étudiants à plein temps. Deux facteurs ont été choisis pour expliquer les taux de chômage par un modèle de transformation de Box-Cox: l'âge et le niveau de scolarité. Nous avons formé des groupes d'âge en décomposant l'intervalle [15,64] en dix groupes, le j -ième groupe étant défini par l'intervalle $[10 + 5j, 14 + 5j]$ pour $j = 1, \dots, 10$ puis en considérant que le milieu de chaque intervalle, $A_j = 12 + 5j$, représente l'âge de tous les membres du groupe correspondant. De même, en ce qui concerne le niveau de scolarité E_k , nous avons formé les groupes en attribuant à chaque personne un nombre d'années de scolarité équivalent au nombre médian, ce qui a donné les six niveaux suivants: 7, 10, 12, 13, 14 et 16. En classant les données en fonction de l'âge et du niveau de scolarité, nous avons obtenu un tableau à double entrée comprenant $I = 60$ estimations d'enquête, P_{jk} , du taux d'emploi P_{jk} . La matrice des covariances estimée V_P reposait sur plus de 450 grappes échantillonnées.

Nous avons considéré le modèle de transformation ci-dessous pour $P_{jk} = F_{jk}(\theta)$ avec effets linéaire et quadratique pour l'âge et effet linéaire seulement pour le niveau de scolarité:

(3.12)
$$v_{jk}(\lambda) = \{F_{jk}/(1 - F_{jk})\}^{(\lambda)}$$
$$= \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, j = 1, \dots, 10, k = 1, \dots, 6.$$

Dans l'équation ci-dessus, $D(\hat{F}_i) = \text{diag}(\hat{F}_i)_{i=1, \dots, I}$, $D(1 - \hat{F}) = \text{diag}(1 - \hat{F}_i)_{i=1, \dots, I}$ et $(\partial F / \partial \theta)'$ est la matrice $I \times (s + 1)$ des dérivées partielles $\partial F_i / \partial \theta_j$ et $\partial F_i / \partial \lambda$ évaluées à θ :

$$\partial F_i / \partial \theta_j = x_{ji} F_i^2 (1 / \bar{Q}_i)^{1+1/\lambda}$$

$$\partial F_i / \partial \lambda = F_i^2 (\bar{Q}_i \log \bar{Q}_i - \bar{Q}_i + 1) \lambda^{-2} (1 / \bar{Q}_i)^{1+1/\lambda}, \quad (3.4)$$

où $\bar{Q}_i = 1 + \lambda \sum_j x_{ji} \theta_j$. Si l'on tient compte du plan de sondage, la matrice des covariances asymptotique estimée de θ , est alors définie (voir Roberts 1985)

$$\text{est cov}(\theta) = (B' \hat{\Delta} B)^{-1} (B' D(\hat{W}) V_p D(\hat{W}) B) (B' \hat{\Delta} B)^{-1}, \quad (3.5)$$

où $\hat{\Delta} = \text{diag}(W_i \hat{F}_i (1 - \hat{F}_i))_{i=1, \dots, I}$ et $D(\hat{W}) = \text{diag}(W_i)_{i=1, \dots, I}$.

Il est aussi intéressant de connaître les erreurs types des résiduels $R_i = \hat{F}_i - F_i$ puisque les résiduels normalisés $R_i / \text{e.t.}(R_i)$ peuvent servir à déceler les proportions de case aberrantes. La matrice des covariances asymptotique estimée du vecteur des résiduels $R = (R_1, \dots, R_I)'$ est définie

$$\text{est cov}(R) = A \text{ est cov}(\theta) A' = V_R, \quad (3.6)$$

où

$$A = I - D(\hat{F}) D(1 - \hat{F}) B (B' \hat{\Delta} B)^{-1} B' D(\hat{W}).$$

La racine carrée des éléments diagonaux, $V_{ii, R}$, de (3.6) correspond à l'erreur type estimée de $R_i, i = 1, \dots, I$.

Correction des tests ordinaires

Le test chi carré ordinaire de validité de l'ajustement du modèle (3.2) est défini par l'expression

$$X^2 = n \sum_{i=1}^I (\hat{F}_i - F_i)^2 W_i / \{F_i (1 - F_i)\} \quad (3.7)$$

tandis que le test du rapport des vraisemblances appliqué dans le même but est défini par

$$G^2 = 2n \sum_{i=1}^I W_i [F_i \log(\hat{F}_i / F_i) + (1 - \hat{F}_i) \log\{(1 - \hat{F}_i) / (1 - F_i)\}], \quad (3.8)$$

où l'expression entre crochets [] est égale à $-\log(1 - \hat{F}_i) - \log F_i$ lorsque $\hat{F}_i = 1$.

Suivant un échantillonnage binomial produit, nous savons que X^2 et G^2 sont identiquement distribuées asymptotiquement selon une loi de chi carré avec $I - s - 1$ d.l., mais cela n'est pas le cas pour des plans de sondage généraux. De fait, X^2 (ou G^2) est distribuée asymptotiquement comme une somme pondérée, $\sum \delta_k W_k$, de variables chi carré indépendantes χ^2_k, W_k , avec, dans chaque cas, un d.l., où les poids δ_k ($k = 1, \dots, I - s - 1$) peuvent être considérés comme des "effets du plan généralisés" (voir Roberts 1985). Suivant un échantillonnage binomial produit, $\delta_k = 1$ pour tous k et $\sum \delta_k W_k$ se ramène à une variable chi carré χ^2 avec $I - s - 1$ d.l. On obtient une correction du premier degré de X^2 (ou de G^2) en considérant $X^2_c = X^2 / \delta$, ou $G^2_c = G^2 / \delta$, comme une variable chi carré χ^2 avec $I - s - 1$ d.l., où

3. MODÈLES DE TRANSFORMATION DE BOX-COX

Les modèles de régression logistique sont largement utilisés pour analyser la variation des proportions estimées qui se rapportent à une variable de réponse binaire. Supposons que la population d'intérêt est répartie en I cases formées selon un ou plusieurs critères de classification. Soit P_i la proportion de réponse de la population dans la case i . Alors, un modèle de régression logistique pour les proportions $P_i = F_i(\beta)$ est défini

$$(3.1) \quad \log\{F_i/(1 - F_i)\} = x_i'\beta, \quad i = 1, \dots, I,$$

où $x_i = (x_{i1}, \dots, x_{is})'$ est un s -vecteur de constantes connues dérivé des niveaux de facteur avec $x_{i1} = 1$, et β est un s -vecteur de paramètres inconnus.

Guerrero et Johnson (1982) ont élargi le champ d'application des modèles de régression logistique en définissant un nouveau paramètre, λ , par l'intermédiaire d'une transformation de Box-Cox des probabilités relatives $F_i/(1 - F_i)$. Leur modèle est défini

$$(3.2) \quad v_i(\lambda) = \{F_i/(1 - F_i)\}^{(\lambda)} = x_i'\beta, \quad i = 1, \dots, I,$$

où β et x_i ont la même définition qu'en (3.1) et

$$\{F_i/(1 - F_i)\}^{(\lambda)} = \begin{cases} \log\{F_i/(1 - F_i)\} & \text{if } \lambda = 0 \\ \lambda^{-1} \{F_i/(1 - F_i)\}^\lambda - 1 & \text{if } \lambda \neq 0. \end{cases}$$

Le modèle de régression logistique (3.1) est un cas particulier du modèle (3.2) (lorsque $\lambda = 0$). Guerrero et Johnson (1982) ont appliqué ce modèle à des données de l'enquête nationale sur les revenus et les dépenses des ménages au Mexique afin d'expliquer la variation du taux d'activité des femmes dans ce pays. Ils ont observé qu'avec une valeur λ de -6.63 , leur modèle donnait un ajustement beaucoup plus précis que le modèle logit ($\lambda = 0$), les valeurs du critère chi carré ordinaire étant de 4.8 (7 d.l.) et de 12.8 (8 d.l.) respectivement. En revanche, ils ont appliqué des méthodes standard pour les proportions binomiales, ne tenant ainsi aucun compte du plan de sondage.

Pseudo EMV

Dans cette section, nous allons étendre au modèle de transformation de puissance (3.2) les méthodes utilisées dans Roberts, Rao et Kumar (1987) pour le modèle de régression logistique. Comme il est difficile de définir des fonctions de vraisemblance pour des plans de sondage généraux, nous allons utiliser des "pseudo-estimateurs du maximum de vraisemblance", $\hat{\beta}$ et $\hat{\lambda}$, que nous allons tirer des équations de vraisemblance de β et λ selon un modèle binomial produit en remplaçant la proportion de réponse simple r_i/n_i par l'estimation d'enquête correspondante P_i de P_i , et en remplaçant n_i/n par l'estimation d'enquête correspondante W_i de la proportion de domaine W_i . Ici, r_i représente le nombre de "succès" dans un échantillon de taille n_i dans la case i , et $n = \sum n_i$. Voir Guerrero et Johnson (1982) pour les équations de vraisemblance selon un modèle binomial produit. D'ailleurs, comme dans Guerrero et Johnson (1982), on peut déterminer les pseudo-estimations les plus vraisemblables (e.m.v.), $\hat{\theta}' = (\hat{\beta}; \hat{\lambda})$, de façon itérative en utilisant une quasi-méthode de Newton. Les proportions de réponse ajustées sont définies par $F = F_i(\theta)$.

Soit V_p matrice des covariances estimée des estimations d'enquête $F = (F_1, \dots, F_I)'$, et

$$(3.3) \quad B = D(F)^{-1}D(1 - F)^{-1}(\partial F/\partial \theta)'.$$

Le tableau 1 donne les estimations de proportions, \hat{p}_{ij} , établies d'après les résultats de la séance d'évaluation de la forme physique; il s'agit d'un tableau croisé mettant en relation la condition physique (bonne = 1, passable = 2, déficiente = 3) et la consommation de tabac (habituelle = 1, occasionnelle = 2, nulle = 3). On peut obtenir la matrice des covariances estimée des \hat{p}_{ij} , \hat{V}_p , en s'adressant aux auteurs.

Comme les deux variables du tableau 1 sont ordinales, nous avons considéré le modèle log-linéaire avec interaction linéaire \times linéaire:

$$\log p_{ij} = \bar{u} + u_{1(i)} + u_{2(j)} + \gamma(v_i - \bar{v})(w_j - \bar{w}), \quad i = 1, 2, 3 \quad j = 1, 2, 3 \quad (2.12)$$

assujetti aux contraintes $\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0$, et selon lequel v_i et w_j sont des scores connus ayant pour moyennes respectives \bar{v} et \bar{w} . Pour des raisons de simplicité, nous avons choisi des scores équidistants: $u_i = 1, 2, 3$; $v_j = 1, 2, 3$. Le modèle (2.12) est de la forme $g(p) = X_0 \beta_0 + X_1 \beta_1$ où $g_{ij}(p) = \log p_{ij}$, $X_0 = K = 1_9$, un vecteur 9×1 de uns, $\beta_0 = \bar{u}$, $\beta_1 = (u_{1(1)}, u_{1(2)}, u_{2(1)}, u_{2(2)}, \gamma)^T$, et

$$X_1' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & 1 & 0 & -1 \\ 1 & -1 & -1 & 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Compte tenu de ce que $H = \text{diag}(\hat{p}_{ij}^{-1}, i = 1, 2, 3; j = 1, 2, 3)$, nous pouvons utiliser le test de Wald défini en (2.6) pour vérifier la validité de l'ajustement du modèle (2.12) en nous servant des proportions \hat{p}_{ij} du tableau 1 et de la matrice des covariances estimée \hat{V}_p . (On peut obtenir copie de cette matrice en s'adressant aux auteurs.) Nous obtenons alors

$$W = 3.59$$

ce qui n'est pas significatif à un seuil de 5 %, étant donné le critère $\chi^2_{7, (0.05)} = \chi^2_7(0.05) = 7.81$ (notons que $T = 9$, $r = 6$). La statistique de Wald W devrait être stable dans cet exemple étant donné le nombre peu élevé de cases T ($= 9$) par rapport au nombre de grappes dans l'échantillon ($= 50$).

Étant donné le modèle (2.12), nous pouvons aussi réaliser un test d'indépendance, c.-à-d. $\gamma = 0$, en utilisant W_1 défini en (2.7), ou W_1^* , fondé sur les estimations lissées β_1^* , et défini en (2.10). Compte tenu de ce que $C_1 = (0, \dots, 0, 1)$, $c_1 = 0$, nous avons

$$W_1 = 8.23, \quad W_1^* = 8.75,$$

ce qui est supérieur à $\chi^2_1(0.01) = 6.63$, la limite supérieure (à 1 %) de la distribution χ^2 avec 1 degré de liberté. L'hypothèse emboîtée de l'indépendance n'est donc pas fondée.

Si nous acceptons le modèle (2.12), nous obtenons les valeurs suivantes pour les estimations par les moindres carrés pondérés, β_1 , et les estimations lissées, β_1^* :

$$\beta_1 = (0.912, -1.550, 0.339, -0.255, -0.086)'$$

$$\beta_1^* = (-2.665, \beta_1^* = (0.917, -1.568, 0.344, -0.262, 0.087)'$$

La valeur estimée β^* peut par ailleurs servir à produire des estimations lissées de p_{ij} , $p_{ij}^* = p_{ij}(\beta^*)$, qui satisfont la contrainte $\sum \sum p_{ij}(\beta^*) = 1$.

avec, comme valeurs initiales, $M_0 = \tilde{M}$, $\tilde{\beta}_0 = (X'NX)^{-1}X'M\tilde{g} = \tilde{\beta}$, $H_0 = H(\tilde{\beta})$ et $p_0 = p(\tilde{\beta})$. De plus, $M_i = (V_{g_i}^* + X_0X_0')^{-1}$ où $V_{g_i}^* = H_i'V^pH_i$, $H_i = H(\beta_i)$ et $p_i = p(\beta_i)$, $i \geq 1$. À la convergence, nous avons $\beta^* = (\beta_0^*, \beta_1^*)'$ comme solution de l'équation suivante:

$$(2.9) \quad X'M(H(\beta)(\beta - p(\beta))) = 0.$$

L'équation (2.9) se ramène à une équation de quasi-vraisemblance (McCullagh, 1983) lorsque V^p est proportionnelle à $V(p)$, une fonction connue de p . La relation de dépendance par rapport à β est mise en évidence dans l'équation ci-dessus si l'on considère que $p = p(\beta)$, $H = H(\beta)$ et $M = V_g^* + X_0X_0' = M(\beta)$. Contrairement à β^* l'estimation lissée K^p de β_1^* et de β_1 sont identiques mais β_1^* peut être plus efficace pour de petits échantillons. Étant donné le modèle (2.1), nous pouvons définir un autre test de Wald de l'hypothèse $C_1\beta_1 = c_1$:

$$(2.10) \quad W_1^* = (C_1\beta_1^* - c_1)' [C_1 \text{est cov}(\beta_1^*) C_1']^{-1} (C_1\beta_1^* - c_1)$$

la variable χ^2 est distribuée asymptotiquement selon une loi de chi carré avec h degrés de liberté et

$$(2.11) \quad \text{est cov}(\beta_1^*) = (X_1^*M^*X_1^*)^{-1},$$

$$\text{et } X_1^* = [I - X_0X_0'M^*]X_1, M^* = (V_g^* + X_0X_0')^{-1} \text{ avec } V_g^* = H^*V^pH^*, \text{ et } H^* = H(\beta^*).$$

Exemple

Les résultats ci-dessus ont été appliqués à un tableau à double entrée tiré de l'Enquête Santé Canada de 1978-1979. Cette enquête avait pour but de recueillir des renseignements précis sur la santé des Canadiens et comprenait deux volets: une interview destinée à tout l'échantillon et une séance d'évaluation de la forme physique destinée à un sous-échantillon. Pour réaliser cette enquête, on a utilisé un plan de sondage à plusieurs degrés, avec stratification et échantillon-nage en grappes, et on a soumis les estimations des totaux ou des proportions par case à une stratification a posteriori selon l'âge et le sexe pour en accroître l'efficacité. Le lecteur est prié de consulter Hidiroglou et Rao (1987) pour une description de l'enquête et des méthodes utilisées pour estimer les effectifs et les proportions par case de même que les variances et les covariances correspondantes. En ce qui concerne l'évaluation de la forme physique, on a estimé la variance au moyen d'une méthode de regroupement de strates puisque, pour quelques-unes des strates, une seule unité primaire d'échantillonnage avait été prélevée.

Tableau 1

Proportions de case estimées dans un tableau 3×3 (Canada):
consommation de tabac \times condition physique (taille de l'échantillon: $n = 2505$)

Consommation de tabac	Condition physique		
	1	2	3
1	0.22005	0.14951	0.16998
2	0.02301	0.00962	0.01146
3	0.20329	0.09933	0.11374

Estimateurs par les moindres carrés pondérés

Nous pouvons exprimer le modèle par la formule

$$(2.1) \quad \hat{g} = g(\beta) = X\beta + \delta$$

où δ est le vecteur d'erreurs avec $P \lim \delta = 0$, et \hat{g} a une matrice des covariances asymptotique singulière $V_g = H'V_pH$, qui a pour estimateur convergent $\hat{V}_g = \hat{H}'\hat{V}_p\hat{H}$, en supposant que \hat{V}_p soit un estimateur convergent de V_p . Dans l'expression précédente, $H = H(\beta)$. Scott, Rao et Thomas (1987) ont défini le meilleur estimateur linéaire asymptotiquement sans biais (MELASB) de β_1

$$(2.2) \quad \hat{\beta}_1 = (X_1'MX_1)^{-1}X_1'\hat{M}\hat{g},$$

ou

$$(2.3) \quad \hat{M} = (V_g + X_0X_0')^{-1}$$

est la pseudo-inverse régulière de V_g , et

$$(2.4) \quad X_1 = [I - X_0X_0'\hat{M}]X_1.$$

Un estimateur convergent de la matrice des covariances asymptotique de $\hat{\beta}_1$ est défini

$$(2.5) \quad \text{est cov}(\hat{\beta}_1) = (X_1'\hat{M}X_1)^{-1}.$$

Tests de Wald

En posant $\hat{\beta} = (X'\hat{M}X)^{-1}X'\hat{M}\hat{g} = (\hat{\beta}_0', \hat{\beta}_1')$, nous pouvons définir un test de Wald pour tester la validité de l'ajustement du modèle (2.1):

$$(2.6) \quad W = (\hat{g} - X\hat{\beta})'\hat{M}(\hat{g} - X\hat{\beta}).$$

La variable χ^2 est distribuée asymptotiquement selon une loi de chi carré avec $T - r$ degrés de liberté (d.l.). Le modèle est considéré comme valide au niveau α si $W > \chi^2_{T-r}(\alpha)$, qui est la borne supérieure $\alpha \chi^2$ de la distribution chi carré avec $T - r$ d.l.

Etant donné le modèle (2.1), nous pouvons définir des tests d'hypothèses linéaires portant sur les paramètres de modèle β_1 . Ainsi, un test de Wald de l'hypothèse linéaire $C_1\beta_1 = c_1$ est défini

$$(2.7) \quad W_1 = (C_1\hat{\beta}_1 - c_1)'[C_1 \text{ est cov}(\hat{\beta}_1)C_1']^{-1}(C_1\hat{\beta}_1 - c_1)$$

la variable χ^2 est distribuée asymptotiquement selon une loi de chi carré avec h degrés de liberté; C_1 est une matrice à rang complet $h \times (r - L)$ de constantes connues ($h > r - L$), et c_1 est un h -vecteur de constantes connues. L'hypothèse est rejetée au niveau α si $W_1 > \chi^2_h(\alpha)$, qui est la borne supérieure $\alpha \chi^2$ de la distribution chi carré avec h degrés de liberté. Notons que β_0 ne doit pas être inclus dans l'hypothèse linéaire puisqu'il est déterminé par les contraintes du plan $K'p = K'g^{-1}(X\hat{g}) = \pi$.

Version lissée du MELASB et tests de Wald correspondants

Nous pouvons aussi obtenir par itération une version lissée du MELASB de β_1 , disons $\hat{\beta}_1^*$:

$$(2.8) \quad \hat{\beta}_{t+1} = \hat{\beta}_t + (X'M_tX)^{-1}X'M_tH_t'(\hat{\beta} - p_t), \quad t = 0, 1, 2, \dots$$

linéaires touchant les probabilités (ou proportions). Nous définissons par la même occasion les tests de validité de l'ajustement et de sous-hypothèses de Wald ainsi qu'une version lissée des estimateurs par les MCP et les tests de sous-hypothèses de Wald correspondants. Ces méthodes doivent être utilisées uniquement lorsque le nombre de cases dans le tableau de contingence est faible ou que le nombre de grappes échantillonnées est relativement élevé.

Dans la section 3, nous étendons les méthodes utilisées pour les modèles de régression logistique aux modèles de Box-Cox, qui renforcent des transformations de puissance des probabilités relatives par case. Comme le démontrent Guerrero et Johnson (1982) dans le contexte de proportions binomiales, les modèles de Box-Cox peuvent donner un ajustement beaucoup plus précis que les modèles de régression logistique, ceux-ci étant un cas particulier de ceux-là.

Dans la section 4, nous décrivons les méthodes qui servent à tester l'égalité des paramètres de deux modèles logit qui correspondent à deux périodes différentes. Si l'hypothèse de l'égalité est acceptée, on peut obtenir pour la période courante des estimations "lissées" des proportions par case plus efficaces que les estimations lissées correspondantes qui reposent uniquement sur les données de la période courante.

Dans la section 5, nous étendons les résultats obtenus avec les modèles de régression logistique à une catégorie de modèles de réponse polytomique. Le modèle de réponse ordonnée de McCullagh (1980) est analysé en détail.

Enfin, dans la section 6, nous décrivons le logiciel utilisé pour appliquer les méthodes ci-dessus.

2. ESTIMATEURS PAR LES MOINDRES CARRÉS PONDÉRÉS ET TESTS DE WALD

La méthode de Koch, Freeman et Freeman (1975) vise à estimer les paramètres de modèles linéaires généralisés de la forme $g^*(p) = X^* \beta^*$, à l'aide d'une estimation d'échantillon, \hat{p} , des probabilités par case pour la population, désignées par un T -vecteur p , et d'une estimation convergente de $\text{cov}(\hat{p}) = V_p$ (par exemple). Selon cette méthode, la matrice des covariances asymptotique du vecteur $g^*(p)$ de dimension u est supposée régulière ($u > T$); ou, de nombreux modèles, y compris le modèle log-linéaire classique, ont la forme $g(p) = X\beta$, où $g(p)$ est un T -vecteur avec une matrice de covariances asymptotique singulière et X est une matrice à rang complet $T \times r$ de constantes connues. Il est possible de transformer ces modèles en forme régulière $g^*(p) = X^* \beta^*$, comme le font Grizzle et Williams (1972) pour le modèle log-linéaire, mais Scott, Rao et Thomas (1987) ont élaboré la méthode uniforme suivante pour les modèles à matrice singulière en ayant recours à la théorie optimale pour les modèles linéaires à matrice de covariances singulière.

Les probabilités par case p et \hat{p} sont soumises à des contraintes linéaires de la forme $K'p = \pi$ et $K'\hat{p} = \pi$, où K est une matrice à rang complet $T \times L$ de constantes connues et π est un L -vecteur de constantes connues π_i ($L > T$). Par conséquent, la matrice des covariances de \hat{p} sera singulière. Par exemple, dans le cas d'un échantillonnage stratifié avec plan de sondage complexe à l'intérieur des strates, nous pouvons écrire $K = I_L \otimes I_m$, $\pi_i = n_i/n$ ($i = 1, \dots, L$) et $p = (p_{11}, \dots, p_{1m}, \dots, p_{L1}, \dots, p_{Lm})'$ où $p_{ij} = (n_i/n)\hat{p}_{ij}$, \hat{p}_{ij} étant la probabilité pour la catégorie j dans la strate i ($\sum_j \hat{p}_{ij} = 1$, $i = 1, \dots, L$; $j = 1, \dots, m$), n_i étant la taille de l'échantillon pour la strate i , $\sum n_i = n$, I_m étant un m -vecteur formé de uns, I_L étant la matrice unité d'ordre L et \otimes désignant la produit tensoriel.

Supposons que l'on puisse exprimer $X\beta$ par la formule $X\beta_0 + X_1\beta_1$, où X_0 est une matrice $T \times L$ telle que $K'H^{-1}X_0$ est régulière et où $H = (\partial g/\partial p)'$ est la matrice $T \times T$ des dérivées partielles de $g(p)$. En particulier, X_0 peut être assimilée à K si la matrice des contraintes K est comprise dans X , comme on le suppose souvent. Comme les restrictions s'appliquent à p supposent des contraintes pour les paramètres β , on peut déduire précisément β_0 des contraintes pour une valeur β_1 donnée.

dans ce sens, surtout en ce qui a trait à l'analyse de données quantitatives recoupées. Cet article portera plus spécialement sur l'analyse de données quantitatives mais soulignons que l'on a obtenu des résultats importants pour d'autres genres d'analyses: analyse de régression (Fuller 1975; Nathan et Holt 1980; Pfefferman et Nathan 1981; Scott et Holt 1982), analyse en composantes principales (Skinner, Holmes et Smith 1986), analyse factorielle (Fuller 1986), régression logistique avec covariables continues (Binder 1983).

Rao et Scott (1984) ont fait une analyse systématique de l'effet du plan de sondage sur le test chi carré de Pearson et le test du rapport des vraisemblances pour des tableaux à plusieurs entrées suivant des modèles log-linéaires hiérarchiques. Ils ont aussi déterminé les corrections du premier degré de tests ordinaires, qui peuvent être calculées à partir de tableaux publiés comprenant les "effets du plan" pour les estimations par case et les totaux marginaux, ce qui facilite les analyses secondaires faites à partir de rapports publiés (voir aussi Gross 1984; Bedrick 1983; Rao et Scott 1987). Ces corrections du premier degré tiennent compte du plan de sondage en ce sens que l'erreur de première espèce réelle de tests fondés sur les statistiques corrigées se rapproche plus du niveau nominal que l'erreur de première espèce de tests ordinaires, qui peut être très élevée. Rao et Scott (1984) ont aussi défini des corrections du second degré, plus précises, fondées sur l'approximation de Satterthwaite d'une somme pondérée de variables χ^2 indépendantes, mais pour effectuer ces tests, il faut connaître la matrice des covariances estimée des estimations par case. Il existe d'autres méthodes qui tiennent compte du plan de sondage, notamment le test de Wald, qui repose sur les moindres carrés pondérés (Koch, Freeman et Freeman 1975), et le test chi carré avec estimateur jackknife (Fay 1985); dans les deux cas, il faut connaître soit la matrice des covariances estimée complète ou des données de grappe. Fay (1985) et Thomas et Rao (1987) ont montré que la statistique de Wald, quoique asymptotiquement juste, peu devenir très instable lorsque le nombre de cases du tableau de contingence augmente et que le nombre de grappes échantillonnées diminue, ce qui a pour effet de porter l'erreur de première espèce à un niveau inacceptable. Par ailleurs, les tests jackknife de Fay et les corrections de Rao-Scott donnent des résultats satisfaisants dans des conditions très générales. Dans certains cas, on peut remédier à l'instabilité de la statistique de Wald en regroupant des cases du tableau selon les vecteurs propres qui correspondent aux valeurs propres non négligeables de la matrice des covariances estimée redressée en fonction de particularités qui découlent de contraintes linéaires touchant les probabilités (voir Singh 1985; Singh et Kumar 1986).

Roberts, Rao et Kumar (1987) ont supposé un modèle de régression logistique pour les proportions par case (domaine) rattachées à une variable de réponse binaire, et ont déterminé les corrections du premier degré du test chi carré ordinaire et du test du rapport des vraisemblances pour la validité de l'ajustement et les hypothèses emboîtées. Ils ont aussi déterminé les bornes supérieures de ces corrections, qui dépendent uniquement des effets du plan des proportions de réponse par case, dans le but de faciliter les analyses secondaires faites à partir de tableaux publiés. Scott (1986) a proposé une méthode par laquelle on applique des tests ordinaires à des données transformées obtenues à partir des données originales et des effets du plan par case. Roberts, Rao et Kumar (1987) ont aussi déterminé les corrections du second degré des tests ordinaires mais rappelons-nous que pour exécuter ces tests, il faut connaître la matrice des covariances estimée des proportions de réponse par case. Ils ont élaboré du même coup des méthodes diagnostiques pour repérer les valeurs aberrantes et les points déterminants tout en tenant compte du plan de sondage.

Cet article a principalement pour but d'exposer des variantes des méthodes mentionnées ci-dessus et d'illustrer l'application de ces méthodes à des données tirées de grandes enquêtes comme l'Enquête Santé Canada (1978-1979) et l'enquête sur la population active du Canada. Dans tout l'exposé, on suppose que l'utilisateur connaît dans son entier la matrice des covariances estimée des estimations par case. Dans la section 2, nous présentons les estimateurs par les moindres carrés pondérés (MCP) des paramètres de modèles linéaires généralisés comportant des matrices de covariances singulières dont l'existence est attribuable à des contraintes

Analyse de données d'enquête avec variables de réponse qualitatives: méthodes et logiciels

J.N.K. RAO, S. KUMAR, et G. ROBERTS¹

RÉSUMÉ

Depuis une dizaine d'années environ, beaucoup de chemin a été parcouru dans l'élaboration de méthodes d'analyse statistique qui tiennent compte de la complexité du plan de sondage. Les progrès les plus notables ont été observés dans l'analyse de données quantitatives recoupées. Mentionnons à ce propos l'estimation par les moindres carrés pondérés de modèles linéaires ou des modèles de régression l'ajustement et de sous-hypothèses de Wald correspondants, la correction de tests chi carré ordinaires ou de tests du rapport des vraisemblances selon des modèles log-linéaires ou des modèles de régression logistique avec variable de réponse binaire, et les tests chi carré avec estimateur jackknife. Cet article vise à décrire l'application de versions élargies de ces méthodes à des données d'enquêtes complexes. Ainsi, la méthode de Scott, Rao et Thomas (1989) pour régression pondérée avec matrices de covariances singulières est appliquée à des données de l'Enquête Santé Canada (1978-1979). Des méthodes pour modèles de régression logistique sont étendues à des modèles de Box-Cox comprenant des transformations de puissance de probabilités relatives par case et sont appliquées à des données de l'enquête sur la population active du Canada. Par ailleurs, nous appliquons à des données de la même enquête des méthodes qui permettent de tester l'égalité des paramètres de deux modèles de régression logistique qui correspondent à deux périodes distinctes. Enfin, nous analysons une catégorie de modèles de réponse polytomique et appliquons des tests chi carré corrigés à des données de l'Enquête Santé Canada (1978-1979). Nous décrivons aussi brièvement le logiciel utilisé dans les circonstances (programmes SAS exécutés sur un gros ordinateur).

MOTS CLÉS: Correction de tests chi carré; régression logistique; transformations de puissance; tests de Wald; moindres carrés pondérés.

1. INTRODUCTION

Les spécialistes des sciences sociales, des sciences de la santé et d'autres disciplines utilisent fréquemment les méthodes statistiques standard qui reposent sur l'hypothèse des observations indépendantes et identiquement distribuées. Ces méthodes se retrouvent aussi dans des logiciels statistiques courants comme SPSSX, BMDP, SAS et GLIM. Dans la pratique toutefois, beaucoup de données proviennent d'enquêtes à plan de sondage complexe où il est question d'échantillonnage en grappes et de stratification, de sorte que l'on risque de faire de fausses inférences si on applique les méthodes standard à ces données sans une correction qui tienne compte du plan de sondage. En particulier, on peut sous-estimer fortement les erreurs types des estimations de paramètres et les intervalles de confiance correspondants si, dans l'analyse des données, on ne tient pas compte de la complexité du plan de sondage. En outre, l'erreur de première espèce réelle des tests d'hypothèses peut être beaucoup plus élevée que le niveau nominal. Les analyses préliminaires de données, par exemple les analyses de résidus visant à déceler les écarts du modèle, sont également touchées. Kish et Frankel (1974) et d'autres ont fait ressortir quelques-uns des problèmes que soulève l'application de méthodes standard et ont fait valoir la nécessité d'élaborer de nouvelles méthodes qui tiennent compte de la complexité du plan de sondage. Au cours des dix dernières années, on a fait des progrès importants

¹ J.N.K. Rao, Département de mathématiques et de statistiques, Université Carleton, Ottawa (Ontario); S. Kumar et G. Roberts, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario).

Dans ce numéro

De plus en plus, les statisticiens découvrent les risques liés à l'utilisation de méthodes statistiques standard pour l'analyse de données provenant d'enquêtes à plan de sondage complexe. La section spéciale du présent numéro contient trois articles de fond sur l'analyse de données qualitatives tirées d'enquêtes à plan de sondage complexe. Tim Holt a contribué à la réalisation de cette section.

L'article de Rao, Kumar et Roberts, qui est le premier article commenté à être publié dans Techniques d'enquête, décrit les progrès qui ont été réalisés dans l'analyse des données qualitatives recoupées, définit des versions élargies des méthodes élaborées dernièrement et applique ces versions aux données de deux grandes enquêtes complexes. De plus, les auteurs traitent brièvement l'aspect informatif. L'article est suivi de commentaires de Fay, de Skinner et de Molina et d'une réponse des auteurs.

Thomas décrit une étude de Monte Carlo au moyen de laquelle il analyse plusieurs méthodes de construction d'intervalles de confiance simultanées pour proportions avec un plan d'échantillonnage en grappes à deux degrés. Il montre que certaines de ces méthodes sont peu efficaces, le niveau de confiance réel s'écartant sensiblement du niveau théorique. En conclusion, il propose des critères pour le choix de la méthode la plus appropriée.

Le dernier article de la section spéciale, rédigé par Morel, porte sur la régression logistique. À l'aide des résultats d'une étude de Monte Carlo, l'auteur montre que, pour de petits échantillons, la méthode du développement de Taylor modifiée produit des biais moins élevés que la méthode delta habituelle pour l'estimation d'une matrice de covariances.

La bibliographie de la méthode des réponses randomisées de Nathan, qui a paru dans un numéro antérieur de Techniques d'enquête, témoigne des nombreuses recherches qui ont été faites sur le sujet. Dans son article, Franklin présente un nouveau modèle de randomisation des réponses pour des populations dichotomiques. Ce modèle est général en ce qu'il prévoit l'utilisation de la randomisation avec distribution continue et des essais multiples pour chaque répondant. L'auteur s'arrête plus spécialement au cas de la randomisation avec distribution normale. MacGibbon et Tomberlin analysent le problème de l'estimation pour petites régions avec plan de sondage complexes. Leur estimateur empirique de Bayes représente un compromis entre l'estimateur classique, qui est non biaisé mais fort variable, et l'estimateur synthétique, plus stable mais susceptible d'être fortement biaisé.

Sunter présente une méthode de mise à jour d'un échantillon PPTSR, qui vise à conserver les mêmes unités primaires d'échantillonnage. Cette méthode diffère de celles proposées antérieurement par Kish et Scott (1971) et Fellegi (1963) en ce qu'elle vaut pour n'importe quelle taille d'échantillon et qu'elle ne nécessite pas une énumération de tous les échantillons possibles. Son utilisation est particulièrement précieuse pour la mise à jour d'échantillons à plusieurs degrés où l'addition de nouvelles UPF est une opération qui peut s'avérer coûteuse. Au Canada, le fichier de Revenu Canada et celui des allocations familiales servent à établir des estimations de la population des provinces dans les années intercensitaires. Verma et Kaby analysent la cohérence des estimations établies à l'aide de ces deux sources. Ils comparent aussi ces estimations aux chiffres du recensement de 1986. Swanson présente une méthode qui permet de construire des intervalles de confiance pour des estimations postcensitaires de la population. Il montre que le test de Wilcoxon peut servir à déterminer si un modèle doit être modifié par suite de changements structurels postcensitaires. L'auteur montre à l'aide de données empiriques que si l'on n'opère pas les changements nécessaires, on obtient des intervalles dont le niveau de confiance est moins élevé que prévu.

TABLE DES MATIÈRES

Dans ce numéro	167
Section spéciale – Analyse de données	

J.N.K. RAO, S. KUMAR et G. ROBERTS	
Analyse de données d'enquête avec variables de réponse qualitatives: méthodes et logiciels	169
Commentaires: R.E. FAY	189
C.J. SKINNER	191
E.A. MOLINA	193
Réponse: Auteurs	195

D.R. THOMAS	
Intervalles de confiance simultanés pour proportions suivant un modèle d'échantillonnage en grappes	197
J.G. MOREL	
Régression logistique selon des plans de sondage complexes	213

L.A. FRANKLIN	
Échantillonnage pour populations dichotomiques par la méthode des réponses randomisées avec randomisation continue	235
B. MacGIBBON et T.J. TOMBERLIN	
Estimation de proportions pour petites régions par des méthodes empiriques de Bayes	247
A. SUNTER	
Mise à jour de la taille de population dans un plan PPTS.R	263
R.B.P. VERMA et R. RABY	
Utilisation des fichiers administratifs pour estimer la population au Canada	271

D.A. SWANSON	
Intervalles de confiance pour les estimations postcensitaires de la population: une étude de cas pour les petites régions	281

Remerciements	291
---------------------	-----

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

M.P. Singh

D. Roy

R. Platek

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *U. of Western Ontario*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

D. Drew, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentileman, *Statistique Canada*

M. Gonzalez, *U.S. Office of*

Management and Budget

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*

Rédacteurs adjoints

J. Gambino, J.-L. Tamby et A. Théberge, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'analyse statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, 4^e étage, Edifice Jean-Jalton, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 30,00\$ par année au Canada, et de 35,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6. Un prix réduit, soit 16,00\$ (E.-U.) (20,00\$ Can.), est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA
DÉCEMBRE 1989

Publication autorisée par
le ministre de l'Expansion industrielle régionale
©Ministre des Approvisionnements
et Services Canada 1990

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable d'une
autorisation écrite du Groupe des programmes et produits
d'édition, agent intérimaire aux permissions, administration
des droits d'auteur de la Couronne, Centre d'édition
du gouvernement du Canada, Ottawa, Canada KIA 0S9.

Mars 1990

Prix: Canada, \$30.00 par année
Autres pays, \$35.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 15, n° 2

ISSN 0714-0045

Ottawa

TECHNIQUES D'ENQUÊTE

VOLUME 15, NUMÉRO 2
DÉCEMBRE 1989

UNE REVUE
DE
STATISTIQUE CANADA

Canada

Statistics
Canada

Statistique
Canada



12-
bal

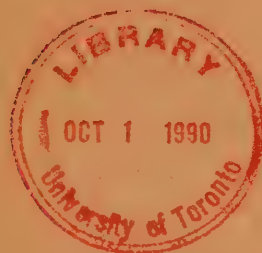


Statistics
Canada

Statistique
Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 16, NUMBER 1
JUNE 1990

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1990

Published under the authority of
the Minister of Industry, Science and Technology

©Minister of Supply
and Services Canada 1990

All rights reserved. No part of this publication may be
reproduced, stored in a retrieval system or transmitted
in any form or by any means, electronic, mechanical,
photocopying, recording or otherwise without
prior written permission of the
Minister of Supply and Services Canada

September 1990

Price: Canada: \$30.00 a year
United States: US\$36.00 a year
Other Countries: US\$42.00 a year

Catalogue 12-001, Vol. 16, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

B. Afonja, <i>United Nations</i>	R.M. Groves, <i>U.S. Bureau of the Census</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	D. Holt, <i>University of Southampton</i>
D. Binder, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
E.B. Dagum, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
J.C. Deville, <i>INSEE</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
W.A. Fuller, <i>Iowa State University</i>	C.E. Särndal, <i>University of Montreal</i>
J.F. Gentleman, <i>Statistics Canada</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

J. Gambino, L. Mach and A. Thériberge, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$30 per year in Canada, US \$36 in the United States, and US \$42 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 16, Number 1, June 1990

CONTENTS

In This Issue	1	
Special Section – History and Emerging Issues in Censuses and Surveys		
J.N.K. RAO and D.R. BELLHOUSE		
History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis.....	3	
Comments: T.M.F. SMITH	26	
S.E. FIENBERG and J.M. TANUR		
A Historical Perspective on the Institutional Bases for Survey Research in the United States	31	
Comments: R.M. GROVES	47	
B.A. BAILAR		
Contributions to Statistical Methodology from the U.S. Federal Government	51	
Comments: G.J. BRACKSTONE	58	
L. KISH		
Rolling Samples and Censuses	63	
Comments: F. SCHEUREN	72	
M.H. HANSEN		
Comments on Articles in the Special Section	81	
REPLIES: J.N.K. RAO and D.R. BELLHOUSE		87
S.E. FIENBERG and J.M. TANUR		89
B.A. BAILAR		91
L. KISH		93
T. DALENIUS and C.-E. SÄRNDAL		
Some Developments of Sampling Techniques and their Use in Official Statistics in Sweden	95	
<hr/>		
P.S. KOTT		
Variance Estimation when a First Phase Area Sample is Restratified.....	99	
D.B. WHITE		
Estimation Using Double Sampling and Dual Stratification	105	
C. JULIEN and F. MARANDA		
Sample Design of the 1988 National Farm Survey	117	

CONTENTS – Concluded

D.A. HAY

Does the Method Matter on Sensitive Survey Topics? 131

E.R. LANGLET

Use of Cluster Analysis for Collapsing Imputation Classes 137

Y. BÉLAND and A. THÉBERGE

An Example of the Use of Randomization Tests in Testing the Census

Questionnaire..... 145

P.J. CANTWELL

Variance Formulae for Composite Estimators in Rotation Designs 153

In This Issue

In this issue's special section, we take a look back and a look forward. Our contributors to this section are well-known survey statisticians who bring a wealth of experience and knowledge. By looking back with clarity to developments in our field, they enable us to look forward to areas of emerging interest. With one exception, each paper has discussants, with a reply by the authors.

Rao and Bellhouse present an historical perspective on sample survey theory and methods. Beginning with a discussion of some of the earliest developments in the field, they then take us through the design-versus model-based debate, variance estimation methods, analysis of survey data and recent developments in computer software. The paper includes an extensive bibliography. Smith's comments complement the paper, providing a somewhat different perspective, including some thoughts on the position of sample survey theory relative to "mainstream" statistics.

Beginning with a discussion of the role of governments and social researchers in the earliest sample surveys and censuses, Fienberg and Tanur describe the institutional bases for survey research, particularly in the United States. Among the organizations considered are government agencies, statistical associations, polling firms and universities. The authors discuss recent developments including increased telephone interviewing and cognitive aspects of surveys. They end by discussing links among the various sectors which make up the field. In his discussion, Groves also looks at the sectors and states that movement of people among them has been less common than Fienberg and Tanur's examples suggest. He also adds substantially to the list of recent developments.

Whereas Fienberg and Tanur look at government institutions as one component out of several, Bailer focuses on the important role played by the U.S. Bureau of the Census in the development of sample survey methods. She discusses the motivation for, and development of, various methods and approaches including sampling and seasonal adjustment. The paper concludes with a look to the future. Brackstone emphasizes that practical problems gave rise to the advances discussed by Bailer. He also adds several other contributions made by Statistics Canada and other agencies to those mentioned by Bailer. Brackstone also points to the importance of a suitable environment to encourage innovation.

Kish discusses alternatives to current periodic censuses. He rekindles the debate on the feasibility of replacing them by rolling censuses. He discusses the use of administrative data in this context, pointing out the existence of good sources of data in some countries. An important issue is how to cumulate data from rolling samples and censuses. Various alternatives are discussed. In his discussion, Scheuren points out that Kish is, in effect, advocating a major shift in our way of thinking – always a difficult task. While Scheuren feels that pure rolling censuses are likely to be too expensive, variations, along with the use of improved administrative data, should be feasible. Both authors agree that there is much research required for further progress.

We are pleased to have Morris Hansen, who participated in many important developments mentioned by the authors, as a discussant of all the above papers. He adds important historical details and corrects some errors and misconceptions. One item of particular interest is Hansen's discussion of the reluctance to introduce sampling – something which we now tend to take for granted. His insightful comments on individual topics are too numerous and varied to summarize here.

Dalenius and Särndal initially intended to discuss Bailer's paper, but their paper metamorphosed into a history of sampling techniques in Sweden. As such, it serves as a summary and update of Dalenius's 1957 book.

The remaining papers in this issue of **Survey Methodology** deal with a diversity of topics. Kott proposes an unbiased estimator of variance for a two-phase sampling design where both phases are stratified simple random sampling. Such designs are commonly used, especially in agricultural surveys.

Two-phase sampling with stratification at both phases is also the subject of White's paper. An estimator due to Vardeman and Meeden which uses prior information is studied via simulation. Some theoretical results are also given for the case where the prior information is not used.

Julien and Maranda describe the sample design used for the National Farm Survey since 1988. The efficiency of the new design is evaluated by comparing the precision of the survey estimates for 1988 to those for 1987, as well as to the expected precision obtained during the development of the new design.

The results of a study in Saskatchewan are analyzed by Hay to examine the effects on responses of the method of data collection: self-administered questionnaire versus personal interview. Although statistically significant differences are found, they are not of sufficient magnitude to be of practical importance.

Langlet studies the use of cluster analysis to deal with the problem of imputation for item nonresponse. This technique would be especially useful in situations where the number of imputation classes is rather large.

Béland and Thériège use randomization tests to compare two questionnaires which were used to study the questions likely to be asked in the 1991 census. Since tests of this type may not be familiar to many survey methodologists, this paper will serve as a useful introduction.

In his paper, Cantwell derives a simple variance expression for a general composite estimator commonly considered for rotating designs. He deals with both single-level and multi-level rotation plans.

The Editor

History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis

J.N.K. RAO and D.R. BELLHOUSE¹

ABSTRACT

Early developments in sampling theory and methods largely concentrated on efficient sampling designs and associated estimation techniques for population totals or means. More recently, the theoretical foundations of survey based estimation have also been critically examined, and formal frameworks for inference on totals or means have emerged. During the past 10 years or so, rapid progress has also been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design. The scope of this paper is restricted to an overview and appraisal of some of these developments.

KEY WORDS: Foundations of inference; Analysis of survey data; Computer software.

1. SOME EARLY MILESTONES

The motivation behind much of the work in survey sampling prior to the 1950's or 60's was the desire to obtain reasonably efficient estimates, at a desired cost, of totals, means, or proportions for large, and increasingly complex-structured, finite populations. A discussion of the early work in sampling human populations may be found in several review papers (see *e.g.*, Hansen, Dalenius and Tepping 1985 and Bellhouse 1988).

The history of the mathematical theory of survey sampling has its origins in the late nineteenth century through the work of the Norwegian statistician A.N. Kiaer. Kiaer was the first to promote what was then called 'the representative method', or sampling, over complete enumeration. What Kiaer (1897) meant by representative sampling was that the sample should mirror the parent finite population. This can be achieved in two ways, by randomization or by balanced sampling through purposive selection. Initially, purposive selection was the preferred method of sample selection, but gradually randomization became a strong competitor to balanced sampling for sample selection. By the 1920's random sampling and purposive selection were both widely used as sample selection techniques. The major theoretical developments in both areas which occurred during this era are summarized in Bowley (1926). This summary includes the development of stratified random sampling with proportional allocation and the derivation of formulae to obtain the precision of an estimate from a purposively selected sample.

The equal footing of random sampling and purposive selection gradually changed after the publication of Neyman's (1934) classic paper. Neyman was able to show, both theoretically and with practical examples, why random sampling was preferable to purposive selection for the large-scale sampling problems of the day. With the publication of the 1934 paper, Neyman also opened up new avenues of development for random sample selection techniques. Previously, Bowley and his followers used only sampling designs with equal inclusion

¹ J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.
D.R. Bellhouse, Department of Statistics, University of Western Ontario, London, Ontario, N6A 5B9.

probabilities for every population unit. Their reasoning was that this method of sampling would provide a representative sample of the universe. Neyman (1934) broke out of this sampling straitjacket with his theories of stratified sampling with “optimal” allocation and cluster sampling with ratio estimation. In both situations, “valid” estimates of population totals, means or proportions are obtained without reliance on a representative sample selected through a design with equal inclusion probabilities. Neyman’s final contribution to the theory of survey sampling is his introduction of cost functions to find the sample allocation in two phase sampling which minimized the variance subject to a fixed budget (Neyman 1938).

Neyman’s fundamental contributions inspired various important extensions of his theory. Among these, we should mention ratio and regression estimation with two-phase sampling (Cochran 1939), determination of “optimal” stratification points and “optimal” allocation with multiple parameters/characters (Dalenius 1957), and sampling on two occasions with partial replacement of units (Jessen 1942) which was subsequently extended by Patterson (1950) and Hansen *et al.* (1953, pp. 470-503) to sampling on more than two occasions (also called rotation sampling). Rotation sampling and associated “composite” estimates are now extensively used to estimate levels and changes from continuing large scale, multi-purpose surveys (*e.g.*, the Current Population Survey (CPS) carried out by the U.S. Bureau of the Census).

Neyman’s work also greatly influenced Morris Hansen, William Hurwitz, and their colleagues at the U.S. Bureau of the Census. Inspired by their practical problems in large-scale survey design and by Neyman’s approach to sampling theory, Hansen and Hurwitz (1943) developed the theory of sampling with probability proportional to size and with replacement (also called PPS sampling). The effect of this approach to multistage surveys is that it provides approximately equal interviewer work loads which makes the administration of a multistage survey easier. This procedure also leads to significant reductions in the variances of the estimates, by controlling the variability arising from unequal cluster sizes without actually stratifying by size and thus allowing stratification on other variables to reduce variance. The theory of Hansen and Hurwitz was extended by Horvitz and Thompson (1952) and Narain (1951) to unequal probability sampling without replacement. By making the inclusion probabilities of units at each stage proportional to their sizes, the desirable features of the Hansen-Hurwitz method are retained, using the so-called Horvitz-Thompson estimator of a population total. The basic work of Horvitz and Thompson and Narain stimulated many theoretical and applied contributions to unequal probability sampling without replacement. Brewer and Hanif (1983) and Chaudhuri and Vos (1988) have provided comprehensive accounts of these developments.

Madow and Madow (1944) have given the basic theory of systematic sampling, and introduced population models to examine the features of systematic sampling. Cochran (1946) introduced the “superpopulation” approach in which the finite population is regarded as being drawn from an infinite superpopulation having certain properties. The expected (or anticipated) variances under the superpopulation model are then compared to study the relative efficiency of alternative sampling strategies. His 1946 paper stimulated much subsequent research in the use of superpopulation models in the choice of sampling strategies and also for model-dependent or model-assisted inference (see Section 2).

Mahalanobis (1946) developed the technique of interpenetrating subsamples, and used it extensively in large-scale surveys in India for assessing both sampling and non-sampling errors. This technique consists of drawing the sample in the form of two or more independent subsamples according to the same sampling scheme such that each subsample provides a valid estimate of the parameter of interest. By assigning the subsamples to different interviewers

(or interviewer teams), a valid estimate of the total variance can be obtained that takes proper account of the correlated response variance component due to interviewers. Deming (1960) used this method (sometimes called replicated sampling) extensively to obtain simple estimates of variance. It has led to resampling techniques such as the jackknife, balanced repeated replication and the bootstrap for getting variance estimates of complex non-linear statistics (see Section 3).

Yet another milestone in the emergence of ideas and theory surrounding complex surveys is the concept of design effect (DEFF), due to Leslie Kish (see Kish 1965, sec. 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance which would be achieved under a simple random sample of the same size. The concept of design effect has been found to be especially useful in the presentation and modelling of sampling errors, and also in the analysis of survey data involving clustering and stratification (see Section 4).

2. THEORETICAL FOUNDATIONS

Although Neyman (1934) and others obtained best linear unbiased estimators for simple designs using the standard Gauss-Markov set-up, the development of traditional sampling theory progressed more or less inductively. Estimators (and designs) which appeared reasonable were considered and their relative properties carefully studied by analytical and/or empirical methods, mainly through comparisons of bias and mean square error, and sometimes also using anticipated mean square error or variance under plausible superpopulation models. As noted by Hansen *et al.* (1983), unbiasedness of estimators under a given design was not insisted on since it "often results in much larger mean square errors than necessary". Instead, asymptotic design consistency of estimators was insisted on, at least when aggregate estimates from reasonably large samples are needed, and the mean square errors of selected asymptotically design consistent estimators were compared to arrive at a suitable estimator (and design). Moreover, in large-scale surveys involving a great many statistics, uniform estimation procedures are often insisted on at the expense of variance inflation for some statistics (compared to alternative estimators tailored to each statistic), due to time, cost and other operational constraints.

Despite the usefulness of the traditional approach, the need for a formal framework for inference from survey data was long felt. Realizing this need, several statisticians have made important contributions to the theoretical foundations of inference from survey data, especially during the past 10-20 years. Several review papers (see *e.g.*, Chaudhuri 1988) and two books (Cassel *et al.*, 1977; Chaudhuri and Vos 1988) discuss various aspects of the theoretical foundations.

Most papers on the theoretical foundations of sampling theory have assumed the following somewhat idealistic set-up. A survey population U consists of N distinct elements identified through the labels $j = 1, \dots, N$. The characteristic of interest y_j (possibly vector-valued) associated with element j can be known **exactly** by observing element j . Thus response or measurement errors are assumed to be absent or ignored if present. The parameter of interest is the population total $Y = y_1 + \dots + y_N$ or the population mean $\bar{Y} = Y/N$ (if N is known). A sample is a subset s of U and the associated y -values, *i.e.*, $\{(i, y_i), i \in s\}$, selected according to a sampling plan which assigns a known probability $p(s)$ to s such that $p(s) \geq 0$ for all $s \in S$ (the set of all possible s) and $\sum_{s \in S} p(s) = 1$. The selection probability $p(s)$ can depend on known design variables $z = (z_1, \dots, z_N)'$, such as stratum indicator variables and size measures of clusters, *i.e.*, $p(s) = p(s | z)$ where z_j is possibly vector-valued. For

probability sampling, the inclusion probabilities $\pi_j = \sum_{\{s: j \in s\}} p(s)$ are positive, which permits unbiased or consistent estimation of Y in the traditional sense. It is also customary to impose the condition that the joint inclusion probabilities $\pi_{ij} = \sum_{\{s: (i,j) \in s\}} p(s)$ be positive, which permits unbiased or consistent variance estimation in the traditional sense.

The basic problem is to make inferences (estimation, variance estimation and constructing confidence intervals), about the total Y by observing a sample selected according to a specified sampling plan $p(s)$ and also using available supplementary data. This involves essentially three steps: (i) choice of a sampling plan; (ii) choice of an estimator \hat{Y} ; (iii) choice of a variance estimator and confidence intervals. There are essentially three different approaches to implement these steps: (i) design-based approach, also called probability sampling approach or randomization approach; (ii) model-dependent approach, also called prediction approach or probability speculation approach (Hájek 1981), (iii) a hybrid approach, called model-based approach or model-assisted approach. Developments to date under each of these three approaches are discussed below.

2.1 Design-based Approach

This approach uses probability sampling both for sample selection and for inference from the data. The probability sampling distribution provides valid inferences irrespective of the population y -values, even in complicated situations, in the sense that the pivotal $t = (\hat{Y} - Y)/s(\hat{Y})$ is approximately $N(0,1)$, at least for large samples, where $s(\hat{Y})$ is the standard error of \hat{Y} . This approach has been criticized on the grounds that such inferences, although assumption-free, refer to repeated sampling from the survey population involving all samples $s \in S$ and the associated probabilities $p(s)$, instead of just the particular s that has been drawn. This criticism can be countered to some extent by using either conditional design-based inference referring to a subset of S that is “relevant” to the particular s or by a model-assisted approach.

Horvitz and Thompson (1952) made a basic contribution to foundational aspects of design-based inference by formulating three classes of linear estimators of Y , and then raising the possibility that the best (minimum variance) estimator among all possible linear unbiased estimators of Y may not exist, even for simple random sampling. Prompted by the Horvitz-Thompson formulation, Godambe (1955) proposed a general class of linear estimators given by $\hat{Y}_b = \sum_{i \in s} b_{si} y_i$, where the weight b_{si} is attached to element i if s is selected and $i \in s$. He proved that no best unbiased estimator of Y could exist in this class, for any sampling plan $p(s)$. Since the criterion of minimum variance had failed, several alternative criteria for the choice of an estimator were proposed. Among these, the admissibility criterion is of some use but is not sufficiently selective in distinguishing between the merits of estimators since too many estimators are admissible. Ghosh (1987) provides an excellent survey of results on admissibility and related criteria in finite population sampling. New criteria that give rise to a unique choice of estimator in the Godambe class for any sampling plan have also been put forth, but the optimality properties established have questionable relevance (see Rao 1971, Rao and Singh 1973). Basu’s (1971) well-known “elephants” example demonstrates the futility of two such criteria, *viz.* necessary bestness and hyperadmissibility.

Godambe (1966) obtained the likelihood function from the sample $\{(i, y_i), i \in s\}$ regarding the N -vector $y = (y_1, \dots, y_N)'$ as the parameter of interest, but it provides no information on $(y_i; i \notin s)$, and hence on the total Y , since the N population units are essentially treated as N separate post strata. A way out of this difficulty is to ignore some of the data to make the sample non-unique and arrive at an informative likelihood function (Hartley and Rao 1968; Royall 1968). Another route is to combine the uninformative likelihood function with exchangeable priors via Bayes theorem to arrive at informative posterior inferences (Ericson 1969).

Conditional inference has attracted considerable attention (and controversy) in classical statistics since Fisher (1925). The choice of a relevant reference set for making conditional inference is not always clear-cut, but in the context of post-stratification it seems sensible to make design-based inferences conditional on the realized strata sample sizes (Durbin 1969). Holt and Smith (1979) provide the most compelling arguments in favour of conditional design-based inference, although their discussion was confined to post-stratification of a simple random sample. Rao (1985) considered a number of real examples involving random sample sizes to illustrate conditional design-based inference and associated difficulties.

Robinson (1987) considered conditional design-based inference from a simple random sample when only the population total X of a concomitant variable x is known. By conditioning on the observed sample mean \bar{x} , he showed that the usual ratio estimator $\hat{Y}_r = (\bar{y}/\bar{x})X$ is conditionally biased. He obtained a conditional bias adjusted ratio estimator given by

$$\hat{Y}_r(adj) = \hat{Y}_r + N(r - b)(\bar{x} - \bar{X})\bar{X}/\bar{x}, \quad (2.1)$$

where $r = \bar{y}/\bar{x}$ and b is the sample regression coefficient. He also showed that a customary variance estimator

$$s_c^2(\hat{Y}_r) = N^2(1 - n/N) \sum_{i \in s} (y_i - rx_i)^2/n(n-1) \quad (2.2)$$

is conditionally biased, while another classical variance estimator

$$s_d^2(\hat{Y}_r) = (\bar{X}/\bar{x})^2 s_c^2(\hat{Y}_r) \quad (2.3)$$

is in fact conditionally unbiased, for large n . Robinson also showed, through a simulation study, that $s_d^2(\hat{Y}_r)$ is very close to the estimator of conditional variance of $\hat{Y}_r(adj)$.

2.2 Model-dependent Approach

A strict model-dependent approach involves purposive sampling, and the model distribution (generated from hypothetical realizations of $\mathbf{y} = (y_1, \dots, y_N)'$ obeying the model) provides valid inferences referring to the particular sample s that has been drawn.

The model-dependent approach was first proposed by Brewer (1963) and extensively studied by Royall and his co-workers, starting with Royall (1970). It is best illustrated under a simple regression model

$$E_m(y_i) = \beta x_i, \quad i = 1, \dots, N; \quad \beta > 0, x_i > 0 \quad (2.4)$$

where E_m denotes the model expectation. It is further assumed that the model variance $V_m(y_i) = \sigma_i^2$ where σ_i^2 is known except for a multiplicative constant, and that the model covariance $\text{cov}_m(y_i, y_j) = 0$, $i \neq j$. Royall (1970) showed that the customary design-unbiased estimator, $N\bar{y}$, under simple random sampling is biased under the model given by (2.4), and that $N\bar{y}$ leads to serious underestimation if the observed sample contains mostly units with small sizes, x_i . These results can also be shown under the conditional design-based approach without assuming a model (Rao 1985).

The best linear model unbiased estimator (or prediction estimator) of Y under the model (2.4) is given by

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in S} \hat{\beta} x_i \quad (2.5)$$

which reduces to the usual ratio estimator \hat{Y}_r if $\sigma_i^2 = \sigma^2 x_i$, where $\bar{s} = U - s$ is the set of non-sampled units and $\hat{\beta}$ is the best linear unbiased estimator of β . The uncertainty in \hat{Y} is measured by $E_m(\hat{Y} - Y)^2 = V_m(\hat{Y} - Y)$ which in the case of \hat{Y}_r reduces to

$$V_m(\hat{Y} - Y) = \{X(X - n\bar{x})/(n\bar{x})\}\sigma^2. \quad (2.6)$$

Since (2.6) decreases as \bar{x} increases, the optimal design is a purposive sample consisting of the n units whose x -values are largest, assuming that the population x_i 's are known. A model unbiased estimator, $s_m^2(\hat{Y} - Y)$, of $V_m(\hat{Y} - Y)$ is obtained from (2.6) by replacing σ^2 with its weighted least squares estimator $\hat{\sigma}^2$, and the resulting pivotal $t_m = (\hat{Y} - Y)/s_m(\hat{Y} - Y)$ is approximately $N(0,1)$ under the model distribution. These theoretical results are impressive, but such model-dependent strategies could lead to serious biases if the assumed model is not completely correct.

To protect against model misspecifications, Royall and Herson (1973 a,b) considered model deviations consisting of second or higher order polynomial terms in x (say q -th order) or an intercept or both, and demonstrated that a balanced sample for which $\bar{x}^{(j)} = \bar{X}^{(j)}$, $j = 1, \dots, q$ provides robustness in the sense that \hat{Y}_r remains model unbiased, where $\bar{x}^{(j)} = \sum_{i \in s} x_i^j/n$ and $\bar{X}^{(j)} = \sum_{i \in U} x_i^j/N$. Further, they have shown that stratification on x with optimal allocation and balanced sampling within each stratum together with the separate ratio estimator of Y provides increased efficiency. Purposively chosen balanced samples have a number of difficulties, nevertheless. First, due to lack of rigorous rules in the sample selection one might be tempted to select units whose x_i are close to \bar{X} (in the case of $q = 1$) which can produce an unrepresentative sample if y is positively correlated with x (Yates 1960, p. 40). Second, balancing is sensitive to departures from the polynomial regression model (Madow 1978, p. 320). Balance is required on the alternative model, which may contain higher-order polynomial terms or other variables or both, and the extra variables in the alternative model must be known in advance. Third, balanced sampling is not feasible for surveys with multiple characters of interest since different samples may be required for each variable.

If the extra concomitant variables z in the model are unknown or unmeasured, Royall and Pfeffermann (1982) recommend simple random sampling since it provides "grounds for confidence that the selected sample is not badly unbalanced on z ", but more recently Royall and Cumberland (1988) seem to favour some form of restricted randomization: "Many techniques, including restricted randomization, stratification and systematic sampling, can be used to help achieve balanced samples. We are not advocating one scheme over another; . . .". In any case, it appears that most advocates of the model-dependent approach seem to recommend probability sampling in some form, as noted by Smith (1984), and hence the main difference between the probability sampling approach and the model-dependent approach is in the choice of the pivotal involving the estimator \hat{Y} and a measure of its uncertainty.

Despite the above-mentioned limitations, the model-dependent approach is useful for studying the conditional performances of conventional procedures, under different plausible models. For instance, the variance estimator $s_a^2(\hat{Y}_r)$ is consistent with the behaviour of the conditional variance $V_m(\hat{Y}_r - Y)$ under the model (2.4) with $\sigma_i^2 = \sigma^2 x_i$, while $s_c^2(\hat{Y}_r)$ is model-biased (Royall and Eberhardt 1975). The variance estimator $s_a^2(\hat{Y}_r)$ is also robust to deviations from the assumption $\sigma_i^2 = \sigma^2 x_i$.

2.3 Model-assisted Approach

Hansen, Madow and Tepping (1983) illustrated the dangers in using model-dependent strategies even when the model is apparently consistent with the sample data. By introducing

a misspecification to the model (2.4) which is not detectable through tests of significance from samples as large as 400, they showed that the design-based coverage of the confidence intervals derived from the model-dependent pivotal $t_r = (\hat{Y}_r - Y)/s_a(\hat{Y}_r)$ is substantially less than the desired level and that it becomes worse as the sample size increases. The poor performance of t_r was due to the asymptotic inconsistency of the estimator \hat{Y}_r with respect to their stratified random sampling design.

The model-assisted approach considers only asymptotically design consistent estimators \hat{Y} that are also model unbiased under an assumed model. Variance estimators that are consistent for the design variance of \hat{Y} and at the same time model unbiased (at least approximately) for the conditional variance $V_m(\hat{Y} - Y)$ are also constructed. Thus the resulting pivotal leads to valid inferences under an assumed model and at the same time protects against model misspecifications in the sense of providing valid design-based inferences irrespective of the population y -values. However, very little attention has been given to studying conditional design-based properties of model-assisted strategies under model misspecifications.

Godambe (1955) assumed the model (2.4) with $V_m(y_i) = \sigma_i^2$ and $\text{cov}_m(y_i, y_j) = 0, i \neq j$, and obtained a lower bound, $\sum_{i \in U} (1/\pi_i - 1)\sigma_i^2$, to the anticipated variance of any design unbiased linear estimator, \hat{Y}_b . He also showed that any fixed sample size plan with $\pi_i = (nx_i)/X$ together with the Horvitz-Thompson estimator, $\hat{Y}_{HT} = \sum_{i \in s} y_i/\pi_i$, attains the lower bound, provided $\sigma_i^2 = \sigma^2 x_i^2$. "Optimal" design unbiased strategies do not exist if $\sigma_i^2 \neq \sigma^2 x_i^2$, and as a result asymptotically optimal strategies were developed by relaxing the restriction to design unbiased estimators and considering asymptotically design-consistent estimators. The generalized regression estimator

$$\hat{Y}_{reg} = \sum_{i \in s} y_i/\pi_i + \hat{\beta} \left(X - \sum_{i \in s} x_i/\pi_i \right) \quad (2.7)$$

for any fixed sample size plan with π_i proportional to σ_i is asymptotically optimal (*i.e.*, the asymptotic anticipated variance attains the lower bound), where $\hat{\beta}$ is a linear model unbiased estimator of β and $E_m E_p (\hat{\beta} - \beta)^2 \rightarrow 0$ as $n \rightarrow \infty$, where E_p denotes the design expectation (Särndal 1980). In particular, the best model unbiased estimator $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i) / (\sum_{i \in s} w_i x_i^2)$ with $w_i = 1/\sigma_i^2$ may be chosen.

If $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i / \pi_i) / (\sum_{i \in s} w_i x_i^2 / \pi_i)$ with $w_i = 1/x_i$ is chosen, then \hat{Y}_{reg} reduces to the simpler form (ratio estimator)

$$\hat{Y}_{reg} = X \hat{\beta} = \sum_{i \in s} g_{si} y_i / \pi_i, \quad (2.8)$$

where $g_{si} = X / (\sum_{i \in s} x_i / \pi_i)$ and g_{si} converges in probability to 1 as $n \rightarrow \infty$ (Särndal and Wright 1984). Särndal, Swensson and Wretman (1989) proposed a new variance estimator for estimators \hat{Y} of the form (2.8) which is design consistent and at the same time approximately unbiased for the conditional variance $V_m(\hat{Y} - Y)$. Their variance estimator for \hat{Y}_{reg} is given by

$$s^2(\hat{Y}_{reg}) = \sum_{i < j \in s} (\pi_i \pi_j - \pi_{ij}) \pi_{ij}^{-1} (g_{si} \tilde{e}_i - g_{sj} \tilde{e}_j)^2 \quad (2.9)$$

where $\tilde{e}_i = (y_i - \hat{\beta} x_i) / \pi_i$. For simple random sampling, $s^2(\hat{Y}_{reg})$ reduces to $s_a^2(\hat{Y}_r)$, given by (2.3), which was justified under the prediction and conditional randomization approaches. Kott (1987) proposed a ratio adjustment to the conventional Yates-Grundy variance estimator, $s_{YG}^2(\hat{Y})$, of any model unbiased asymptotically design consistent estimator \hat{Y} . His variance estimator

$$\hat{s}_{YG}^2(\hat{Y}) = s_{YG}^2(\hat{Y}) [V_m(\hat{Y} - Y)/E_m s_{YG}^2(\hat{Y})] \quad (2.10)$$

is model unbiased and at the same time asymptotically design consistent. However, for estimators of the form (2.8) Särndal *et al.* variance estimator appears simpler since it is obtained simply from the conventional variance estimator $s_{YG}^2(\hat{Y})$ by changing \tilde{e}_i to $g_{si}\tilde{e}_i$.

The conventional regression estimator is obtained by first considering a fixed constant B in place of $\hat{\beta}$ in (2.7), and then substituting a consistent estimator of B_{opt} , the value of B minimizing the design variance. This estimator does not depend on the validity of any model. However, the optimal design variance can be approximately attained in the model-assisted framework by modifying the model (2.4) to $E(y_i) = \beta x_i + \gamma \pi_i$ and then using $(\tilde{\beta}, \tilde{\gamma})'$, the weighted regression estimator of $(\beta, \gamma)'$ with weights $w_i = 1/\pi_i^2$. The resulting estimator of Y reduces to (2.7) with $\hat{\beta}$ changed to $\tilde{\beta}$ (Isaki and Fuller 1982; Montanari 1987). Any other choice of $\hat{\beta}$ in (2.7) will give a larger asymptotic design variance.

Little (1983) argued that only models that yield asymptotically design consistent, best linear model unbiased estimators should be used since the latter estimators are optimal if the model is in fact true. One way to accomplish this is by introducing an additional auxiliary variable $u_i = \sigma_i^2(1 - \pi_i)/\pi_i$ into the model (2.4), *i.e.* by using $E(y_i) = \beta x_i + \gamma u_i$ (Särndal and Wright 1984). If we change the model to $E(y_i) = \beta x_i + \gamma \sigma_i^2/\pi_i + \delta \sigma_i^2$ by adding two auxiliary variables σ_i^2/π_i and σ_i^2 to the model (2.4), then we get an asymptotically design consistent, best linear model unbiased estimator of the form $\hat{Y} = \sum_{i \in s} g_{si} y_i / \pi_i$ (Särndal and Wright 1984). The lower bound to asymptotic anticipated variance is also attained if we choose a sampling plan with π_i proportional to σ_i . The above desirable properties, however, are obtained at the expense of a slight increase in the model variance under the original model (2.4).

Godambe and Thompson (1986) employed the theory of estimating functions to derive design consistent estimators through an assumed model. For example, if y_i is expected to be unrelated to π_i for some character y in a multisubject survey, then the “optimal” estimating function gives the Hájek (1971) estimator of \bar{Y} :

$$\hat{\bar{Y}}_H = \left(\sum_{i \in s} y_i / \pi_i \right) / \left(\sum_{i \in s} 1 / \pi_i \right). \quad (2.11)$$

The superpopulation model here is given by $y_i = \theta + \epsilon_i$, with independent errors ϵ_i , which reflects the situation at hand. The estimator $\hat{\bar{Y}}_H$ avoids the difficulties associated with the Horvitz-Thompson estimator \hat{Y}_{HT}/N , as illustrated by the “elephants” example of Basu (1971). The method of estimating functions looks promising, but further work remains to be done on its use in getting “better” estimators or pivots or both. It is interesting to note that the well-known Fieller method of computing confidence limits for a ratio (Fieller 1932) and the method of Woodruff (1952) for computing confidence limits for medians are essentially equivalent to the method of estimating functions.

The results in Sections 2.2 and 2.3 use models appropriate to unistage sampling. In the case of multistage sampling, the models are more complex due to intra-cluster correlations (Scott and Smith 1969; Montanari 1987). The resulting best linear model unbiased estimators or prediction estimators involve weighted combinations of estimators, where the weights depend on intra-cluster correlations which can be estimated from the sample data. Bellhouse and Rao (1986) investigated the relative efficiency of such estimators, under the repeated sampling framework. Their empirical results suggest that the prediction estimators may not be significantly more efficient than the customary estimator in two-stage sampling with PPS sampling of clusters and simple random sampling within sampled clusters.

If the clusters are regarded as strata and if the strata means are the parameters of interest as in small area estimation, then the prediction estimators of strata means are likely to be significantly more efficient than the customary design-based estimators since the prediction estimators “borrow strength” from all the strata unlike the customary estimators. In the case of two-stage sampling with cluster means as parameters of interest, only a prediction estimator for the nonsampled clusters can be implemented.

3. VARIANCE ESTIMATION AND CONFIDENCE INTERVALS

3.1 Linear Statistics

A substantial part of traditional sampling theory is devoted to the derivation of mean square errors or variances of linear estimators of a total Y , and their estimators. Rao (1979) developed a unified approach for estimators belonging to Godambe’s general linear class, $\hat{Y}_b = \sum_{i \in s} b_{is} y_i$, which enables the derivation of mean square error in a straightforward fashion, and also exhibits the necessary form of any non-negative quadratic unbiased estimator of the mean square error. For multistage designs, a general estimator of Y is of the form $\hat{Y}_{bm} = \sum_{i \in s} b_{is} \hat{Y}_i$, where s now denotes a sample of primary sampling units (psu’s) and \hat{Y}_i is an unbiased linear estimator of psu total Y_i based on subsampling the psu. Unified variance formulae for multistage designs have been worked out by Raj (1966) and Rao (1975).

Large scale surveys often employ many strata, L , with relatively few psu’s n_h , sampled within each stratum h . In fact, it is a common practice to select $n_h = 2$ psu’s within each stratum to permit maximum degree of stratification of psu’s consistent with the provision of a valid variance estimator. If the psu’s are sampled with replacement with probabilities p_{hi} in stratum h , then the estimator of total Y is given by $\hat{Y} = \sum_h \bar{r}_h$, and an unbiased variance estimator is simply obtained as

$$s^2(\hat{Y}) = \sum_h \left\{ \sum_i (r_{hi} - \bar{r}_h)^2 / [n_h(n_h - 1)] \right\}, \quad (3.1)$$

where $\bar{r}_h = \sum_i r_{hi} / n_h$, $r_{hi} = \hat{Y}_{hi} / p_{hi}$ and \hat{Y}_{hi} is an unbiased estimator of the i -th psu total in stratum h ($i = 1, \dots, n_h$; $h = 1, \dots, L$). This stratified design is frequently used in comparing methods for nonlinear statistics (Section 3.2). Because of its simplicity, $s^2(\hat{Y})$ is often used even when the psu’s are sampled without replacement. This procedure leads to overestimation of variance, but the relative bias would be small if the first stage sampling fraction is small.

3.2 Non-linear Statistics

Many non-linear, finite population parameters of interest, θ , such as ratio, regression and correlation coefficients, can be expressed as smooth functions, $g(\mathbf{Y})$ of totals $\mathbf{Y} = (Y_1, \dots, Y_q)'$ of suitably defined variates such that $g(\mathbf{Y}) \propto g_1(Y_1/M, \dots, Y_{q-1}/M)$, where $Y_q = M$, the population size. The parameter θ is estimated by $g(\hat{\mathbf{Y}}) \propto g_1(\hat{Y}_1/\hat{M}, \dots, \hat{Y}_{q-1}/\hat{M})$. Such estimators are well-behaved even when the variates attached to the elements t are not related to the inclusion probabilities π_t ($t = 1, \dots, M$) since $g(\hat{\mathbf{Y}})$ is a function only of the Hájek-type estimators $\hat{Y}_j = \hat{Y}_j/\hat{M}$ of the means \bar{Y}_j . As an example of $g(\hat{\mathbf{Y}})$, the estimator of a finite population regression coefficient $B = \sum (x_t - \bar{X})(y_t - \bar{Y}) / \sum (x_t - \bar{X})^2$ can be written as

$$\hat{B} = [\hat{Z}/\hat{M} - (\hat{X}/\hat{M})(\hat{Y}/\hat{M})][\hat{W}/\hat{M} - (\hat{X}/\hat{M})^2]^{-1}, \quad (3.2)$$

where \hat{X} , \hat{Z} and \hat{W} are the estimators of the totals X , Z and W of the variates x_t , $z_t = y_t x_t$ and $w_t = x_t^2$ respectively.

Variance estimation methods for non-linear statistics, $g(\hat{\mathbf{Y}})$, include the well-known linearization method and resampling techniques like the jackknife, balanced repeated replication (BRR) and the bootstrap. The linearization method is applicable to general sampling designs, but it involves a separate variance formula for each statistic. On the other hand, resampling methods use a single variance formula for all statistics. The jackknife and BRR, however, are strictly applicable only to those designs in which the psu's are sampled **with** replacement (or the first-stage sampling fractions are negligible). The bootstrap seems to be more generally applicable, but it is computationally more cumbersome and its properties have not yet been fully examined.

Linearization method

If we denote the variance estimator of $\hat{Y} = \hat{Y}(y_t)$ for a general design as $v(y_t)$, the linearization method provides a variance estimator for a nonlinear statistic $\hat{\theta}$ as $v(z_t)$ for a suitably defined synthetic variable z_t which depends on the form of $\hat{\theta}$. For a general statistic $\hat{\theta} = g(\hat{\mathbf{Y}})$, the variance estimator is given by

$$s_L^2(\hat{\theta}) = v(z_t) \quad \text{with} \quad z_t = \sum_i y_{ti} g_i(\hat{\mathbf{Y}}), \quad (3.3)$$

(Woodruff 1971), where y_{ti} is the value of i th character for t th unit, and $g_i(\hat{\mathbf{Y}})$ is the partial derivative $\partial g(\mathbf{Y})/\partial Y_i$ evaluated at $\mathbf{Y} = \hat{\mathbf{Y}}$ ($i = 1, \dots, q$). One drawback of the formula (3.3) is that the evaluation of partial derivatives may be difficult in some cases, although useful approximations to the desired partial derivatives can be obtained using numerical methods (Woodruff and Causey 1976). The variance estimator can also be obtained in many cases, without actually evaluating the partial derivatives g_i , by recasting $\hat{\theta}$ as a ratio-type statistic and using the usual variance formula for a ratio. For example, the sample regression coefficient \hat{B} may be expressed as $\hat{B} = \hat{Y}(z_{1t})/\hat{Y}(z_{2t})$ with $z_{1t} = (y_t - \hat{Y})(x_t - \hat{X})$ and $z_{2t} = (x_t - \hat{X})^2$, so that

$$s_L^2(\hat{B}) = v(z_{1t} - \hat{B}z_{2t})/[\hat{Y}(z_{2t})]^2. \quad (3.4)$$

Similar techniques can be used for other statistics like the multiple regression coefficients (Fuller 1975; Folsom 1974). Binder (1983) extended the scope of linearization method to statistics defined implicitly as the solution of a set of nonlinear equations. His formulation covers finite population parameters derived from generalized linear models which include the linear regression model and the logistic regression model.

Resampling methods

We now turn to resampling methods for the commonly used stratified multistage design of Section 3.1. Letting $\hat{\theta}^{hi}$ be the estimator of θ computed from the sample $\{\mathbf{r}_{hi}\}$ after omitting $\mathbf{r}_{hi} = \hat{\mathbf{Y}}_{hi}/p_{hi}$, a jackknife variance estimator of $\hat{\theta} = g(\sum \bar{\mathbf{r}}_h)$ is given by

$$s_J^2(\hat{\theta}) = \sum_h \{(n_h - 1)/n_h\} \sum_i (\hat{\theta}^{hi} - \hat{\theta})^2. \quad (3.5)$$

Several variations of (3.5) can be obtained; for instance, $\hat{\theta}$ in (3.5) may be replaced by $\hat{\theta}^h = \sum_i \hat{\theta}^{hi}/n_h$.

McCarthy (1969) proposed the BRR method for the important special case of $n_h = 2$. A set of J "balanced" half-samples is formed by deleting one psu in the sample from each stratum. This set may be constructed from Hadamard matrices. The BRR variance estimator is given by

$$s_{\text{BRR}}^2(\hat{\theta}) = \sum_j (\hat{\theta}^{(j)} - \hat{\theta})^2 / J, \quad (3.6)$$

where $\hat{\theta}^{(j)}$ is the estimator computed from the j -th half sample. Again, several variations of (3.6) can be obtained. The BRR method has been extended recently to the general case of unequal n_h , using asymmetrical orthogonal arrays (Gupta and Nigam 1987; Wang and Wu 1988).

The bootstrap method for the stratified design involved the following steps (Rao and Wu 1988): (i) Draw a simple random sample $\{\mathbf{r}_{hi}^*\}_{i=1}^{m_h}$ of size m_h with replacement from $\{\mathbf{r}_{hi}\}_{i=1}^{n_h}$, independently for each h . Calculate

$$\bar{\mathbf{r}}_{hi} = \bar{\mathbf{r}}_h + [m_h / (n_h - 1)]^{1/2} (\mathbf{r}_{hi}^* - \bar{\mathbf{r}}_h), \quad \bar{\mathbf{r}}_h = n_h^{-1} \sum_i \mathbf{r}_{hi}$$

and $\tilde{\theta} = g(\sum \bar{\mathbf{r}}_h)$. (ii) Independently replicate step (i) a large number, B , of times and calculate the corresponding estimators $\tilde{\theta}^1, \dots, \tilde{\theta}^B$. (iii) The bootstrap variance estimator of $\hat{\theta}$ is given by

$$s_{\text{BOOT}}^2(\hat{\theta}) = \sum_b (\tilde{\theta}^b - \hat{\theta})^2 / (B - 1). \quad (3.7)$$

Confidence intervals can also be obtained by approximating the distribution of $t = (\hat{\theta} - \theta) / s_J(\hat{\theta})$ by its bootstrap counterpart $\tilde{t} = (\tilde{\theta} - \hat{\theta}) / s_J^*(\tilde{\theta})$, where $s_J^*(\tilde{\theta})$ is obtained from $s_J^2(\hat{\theta})$ by jackknifing the particular bootstrap sample $\{\mathbf{r}_{hi}^*\}$. Two-sided $1 - \alpha$ level "bootstrap- t " confidence intervals on θ are then given by

$$\{\hat{\theta} - \tilde{t}_{\text{UP}S_J}(\hat{\theta}), \hat{\theta} - \tilde{t}_{\text{LOW}S_J}(\hat{\theta})\}, \quad (3.8)$$

where \tilde{t}_{LOW} and \tilde{t}_{UP} are the lower and upper $\alpha/2$ points of \tilde{t} obtained from the bootstrap histogram of $\tilde{t}^1, \dots, \tilde{t}^B$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Also, one could use the linearization variance estimator instead of the jackknife variance estimator in constructing the confidence intervals. For confidence intervals we need a much larger number, B , of bootstrap samples than for variance estimation. Regarding the choice of bootstrap sample sizes m_h , the choice $m_h = n_h - 1$ is attractive since it gives $\bar{\mathbf{r}}_{hi} = \mathbf{r}_{hi}^*$.

Comparison of the methods

Theoretical properties of the methods reported in the literature include the following: (1) All the variance estimators reduce to the "standard" one, $s^2(\hat{Y})$ given by (3.1), in the linear case $g(\mathbf{Y}) = Y$. (2) For smooth functions $g(\mathbf{Y})$, all the variance estimators are asymptotically design consistent (Krewski and Rao 1981). The jackknife variance estimator, however, is known to be inconsistent for nonsmooth functions like the quantiles, even in the case of simple random sampling. Hence, caution should be exercised in using jackknife software. (3) If $n_h = 2$ for all h , then the jackknife and linearization variance estimators are asymptotically equal to high order terms for smooth functions $g(\mathbf{Y})$, indicating that the choice between

these methods in this important special case should depend more on other considerations like computational costs (Rao and Wu 1985). Turning to empirical studies, Kish and Frankel (1974) studied the linearization, jackknife and BRR methods, using data from the Current Population Survey and sample designs with $n_h = 2$ clusters from each of $L = 6, 12$ and 30 strata. They evaluated the empirical coverage probability of the $1 - \alpha$ level confidence intervals, $\hat{\theta} \pm t_{\alpha/2} s(\hat{\theta})$, for ratios, regression and correlation coefficients, where $t_{\alpha/2}$ is the upper $\alpha/2$ -point of a t -variable with L degrees of freedom and $s^2(\hat{\theta})$ is anyone of the variance estimators. The BRR method performed consistently better, in terms of coverage probability, than the jackknife which in turn was better than the linearization method; the observed differences were small for ratios. The methods performed in the reverse order with regard to stability of variance estimator. Other empirical studies in the literature reported similar results. Regarding the bootstrap, a simulation study by Kovar, Rao and Wu (1988) indicates that the bootstrap t -intervals track the nominal error rate in each tail better than the intervals based on the normal approximation to $t = (\hat{\theta} - \theta)/s(\hat{\theta})$, but the bootstrap variance estimators are less stable than those based on the linearization or the jackknife. The second order equivalence of the latter two variance estimators for the special case $n_h = 2$ is also confirmed.

Computationally simpler methods of variance estimation than the previous methods have also been proposed in the literature, *e.g.*, random group method and partially balanced repeated replication, but these variance estimators do not reduce to the “standard” one in the linear case. Methods of constructing models from which sampling errors can be imputed have also been proposed. Such methods are useful in producing “smoothed” standard errors for estimators for which direct computations have not been made, and also in presenting standard errors in a concise form (*e.g.*, graphs) in published reports.

Wolter’s (1985) book gives an excellent introduction to recent developments in variance estimation, and illustrates the methods on data from a variety of large-scale surveys. Recent review papers on variance estimation include Rust (1985) and Rao (1988).

4. ANALYSIS OF SURVEY DATA

Standard methods of data analysis are, in general, based on the assumption of simple random sampling. These methods have also been implemented in standard statistical packages, including SPSS^X, BMDP and SAS. Application of standard methods to survey data without some adjustment for survey design, however, can lead to erroneous inferences, since most such data are obtained from complex sample surveys involving clustering, stratification and unequal probability sampling, and as a result do not satisfy the assumption of simple random sampling. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the effect of design is ignored in the analysis of data. Similarly, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods and emphasized the need for new methods that take proper account of the complexity of survey data. During the past 10 years or so, rapid progress has been made in developing such methods for the following types of analyses: (a) analysis of multi-way contingency tables; (b) analysis of domain means or domain proportions; (c) linear regression analysis; (d) multivariate analysis including principal component analysis and factor analysis. A brief account of some of these developments is given in this section, and the reader is referred to review articles by Nathan (1988), Rao (1987) and Smith (1984), and a book edited by C.J. Skinner, D. Holt and T.M.F. Smith (1989).

4.1 Analysis of Multi-way Contingency Tables

Chi-squared tests (or likelihood ratio tests) are frequently used for the evaluation and selection of parsimonious models on \mathbf{p} , the population cell probabilities, in a multi-way contingency table with T cells. For this purpose, loglinear models are convenient because of their close similarity to analysis of variance in systematically providing test statistics of various hypotheses associated with a multi-way table. Rao and Scott (1984) made a systematic study of the impact of survey design on the standard chi-squared test of goodness-of-fit of a loglinear model, denoted by X^2 . They showed that X^2 is asymptotically distributed as a weighted sum, $\sum \delta_i W_i$, of $T - r - 1$ independent χ^2_1 variables W_i , where the weights δ_i are the eigenvalues of a "generalized design effects" matrix and $T - r - 1$ is the degrees of freedom. This general result shows that the survey design can have a substantial impact on the type I error rate of X^2 . For instance, under a constant design effects clustering model, $\delta_i = \lambda$ for all i , the actual type I error rate, for nominal level α , is approximately given by $Pr[\chi^2_{T-r-1} > \lambda^{-1} \chi^2_{T-r-1}(\alpha)]$ which increases with the clustering effect, λ .

Rao and Scott (1984,7) obtained simple first-order corrections to X^2 which can be computed from published tables that include estimates of design effects (or standard errors) for cell estimates $\hat{\mathbf{p}}$ and their marginal totals, thus facilitating secondary analyses (see also Fellegi 1980, Gross 1984, and Bedrick 1983). A first-order correction refers $X^2/\hat{\delta}$ to χ^2_{T-r-1} , where $\hat{\delta}$ is an estimate of the average design effect $\delta = \sum \delta_i / (T - r - 1)$ or an estimate of an upper bound on δ . The corrected test is asymptotically valid in the case of constant design effects clustering, and in general it should perform well when the variability of the δ_i 's is small. More accurate, second-order corrections that take account of the variability in the δ_i 's can also be obtained by using the Satterthwaite approximation to the weighted sum of independent χ^2 variables (Rao and Scott 1984). These tests, however, require the knowledge of a full estimated covariance matrix of $\hat{\mathbf{p}}$. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975) and the jackknife chi-squared tests (Fay 1985). The latter tests are applicable to survey designs permitting the use of a replication method, such as the jackknife or the BRR. The Wald tests require the full estimated covariance matrix of $\hat{\mathbf{p}}$, whereas the jackknife tests require access to cluster-level estimates.

Fay (1985) and Thomas and Rao (1987) showed that the Wald test which refers to χ^2_{T-r-1} , although asymptotically correct, can become highly unstable as the number of cells in the multi-way table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level, α . On the other hand, Fay's jackknife tests and the Rao-Scott corrections performed well under quite general conditions. A simple modification to the Wald test which refers to an F distribution on $T - r - 1$ and $f - T + r + 2$ degrees of freedom performed better than the Wald test in controlling the type I error rate, where f is the degrees of freedom for estimating the covariance matrix of $\hat{\mathbf{p}}$.

4.2 Analysis of Domain Means or Domain Proportions

Analysis of domain (or subpopulation) proportions associated with a binary response variable is of considerable interest to researchers in social and health sciences, and other subject matter areas. Logistic regression models are extensively used for this purpose in conjunction with standard statistical methods for binomial proportions. Rao and Scott (1987) obtained simple first-order corrections to standard chi-squared tests of goodness-of-fit and of nested hypotheses which can be computed from published tables that include estimates of design effects (or standard errors) of domain proportions. Roberts, Rao and Kumar (1987) derived more

accurate second-order corrections to standard tests, but these require access to a full estimated covariance matrix of domain proportions. Diagnostics for detecting outlying domain proportions and influential points in the factor space were developed as well, again taking the sampling design into account.

Koch, Freeman and Freeman (1975) used weighted least squares methods to analyze domain means of a quantitative variable, y , and developed Wald tests of goodness-of-fit of the model and of linear hypotheses on the model parameters. The performance of Wald tests can be improved, as in Section 4.1, by using an F -modification.

4.3 Linear Regression Analysis

In Section 3.2, we considered design-based inferences on nonlinear, finite population parameters such as the finite population simple regression coefficient B . The pivotal $t = (\hat{B} - B)/s(\hat{B})$ is approximately $N(0,1)$, where \hat{B} is the design-consistent estimator, (3.2), of B , and its standard error, $s(\hat{B})$, can be obtained either through the linearization method as in (3.4) or by using one of the replication methods. This approach readily extends to multiple regression coefficients. The design-weighted estimator \hat{B} or its multiple regression analogue can be obtained by the weighted regression option of standard packages by using the survey weights attached to the sample elements as the weights in the regression. However, the standard error of \hat{B} resulting from this routine remains incorrect.

Some people argue that most users are concerned with inferences on parameters of an appropriate superpopulation model rather than inferences on finite population parameters like B . However, the interest in B can also be justified by considering it as the least squares estimator of the superpopulation parameter β in the model

$$y_i = \alpha + \beta x_i + \epsilon_i \text{ with } E_m(\epsilon_i) = 0, \quad i = 1, \dots, N. \quad (4.1)$$

If the population size is large, then estimating B is effectively equivalent to estimating β , while if the model (4.1) is misspecified to the extent of making β meaningless, then B may still be of interest as the slope of the least squares line fitted to the N -pairs (y_i, x_i) (Godambe and Thompson 1986).

Scott and Holt (1982) used a model-dependent approach to investigate the effect of two-stage sampling on standard regression analysis. They assumed a regression model of the form (4.1) with equi-correlated error terms ϵ_i within each cluster, as in Fuller (1975). This model also holds for the sample pairs (y_i, x_i) , $i \in s$, if the selection probabilities are not related to the dependent variable, as in the case of two-stage random sampling. The results of Scott and Holt indicate that the effect of a positive intra-cluster correlation is to understate the standard errors of parameter estimates, and consequently inflate the type I error rates of customary tests. Wu, Holt and Holmes (1988) made a systematic study of the effect of two-stage sampling on the customary F -statistic, and proposed a correction for the F test for unknown intra-cluster correlation, as an alternative to iterative generalized least squares (GLS) procedure. Both the GLS procedure and the F -correction require known cluster labels which may not be available when the survey data are used for secondary analysis.

If the regression model includes all the design variables z related to the dependent variable, such as stratum indicator variables and size measures of units, and the errors ϵ_i are independent with a constant variance σ^2 , then standard regression analysis is valid under the model-dependent approach (Pfefferman and Smith 1985). However, such models may involve too many parameters to be useful. Also, the design variables may not be of intrinsic interest to the user, or may not be available in secondary analysis. In such situations, we are often interested

in models of the form (4.1), where x is not a design variable. The sample pairs (y_i, x_i) , $i \in s$ however, may not satisfy the model due to sample selection bias. Nathan and Holt (1980) proposed an adjusted regression approach to take account of selection bias, and compared it with ordinary least squares and the design based approach based on \hat{B} and $s(\hat{B})$. This approach assumes specific relationships between the regression variables and the design variables. Their empirical results indicate that ordinary least squares inferences can be highly unreliable, that the design-based approach is basically reliable except under extreme selection schemes, and that the adjusted regression approach performs well. Pfefferman and Holmes (1985) study the robustness of these procedures to misspecification of relationships between the regression variables, and conclude that the adjusted regression approach is very sensitive to model misspecification. The design-weighted estimator \hat{B} is robust, but a more efficient estimator is obtained by modifying the adjusted regression estimator to be design-consistent for the finite population regression coefficient, B .

4.4 Multivariate Analysis

The methods in Section 4.2 for the analysis of domain means can be extended to the multivariate case of domain mean vectors, but no detailed studies of such extensions have been reported in the literature. The literature on multivariate analysis of survey data is largely devoted to the analysis of covariance structures, in particular to principal component analysis and factor analysis. Bebbington and Smith (1977), Tortora (1980) and Skinner, Holmes and Smith (1986) investigated the effect of sample design on standard principal component analysis. Their results indicate that the application of standard methods, without some adjustment for the sample design, can lead to erroneous inferences. In particular, the estimators of eigenvalues and eigenvectors of the covariance matrix, Σ_y , can be severely biased for non-self-weighting sample designs. Skinner, Holmes and Smith (1986) proposed maximum likelihood (ML) estimators, under a multivariate normal model, and probability-weighted (or design-based) estimators, to adjust for the effects of the sample design. Their simulation study indicates that both estimators perform well unconditionally, while the probability-weighted estimators exhibit a conditional model bias. The ML estimators, however, may be sensitive to model misspecification. A probability-weighted version of the ML estimators may be more robust, as demonstrated by Pfefferman and Holmes (1985) in the context of the adjusted regression approach (section 4.3). Fuller (1987) derived design-based estimators of the parameters in factor analysis, and the estimated covariance matrix of the estimators. He showed that the estimated variances based on normal theory can seriously underestimate the true variances of the factor estimators.

5. COMPUTER SOFTWARE

Several computer package programs for variance estimation in complex surveys were developed in the mid to late 1970's, often in conjunction with programs for regression analysis of survey data. Wolter (1985, pp. 393-412) reviewed the latest versions of these programs to about 1985. Among the programs listed by Wolter, the ones most commonly used are CLUSTERS (Verma and Pearce 1977), the programs &PSALMS and &REPERR in the OSIRIS IV system (Vinter 1980 and Lepkowski 1982), SUDAAN (Shah 1981a, 1981b, 1982 and Holt 1979), HESBRR (Jones 1983) and SUPER CARP (Hidiroglou, Fuller and Hickman 1980). The programs HESBRR and the OSIRIS IV program &REPERR use balanced repeated replication as the variance estimation technique; the remaining three use the Taylor linearization method.

Cohen, Burt and Jones (1986) evaluated the variance estimation programs for means and ratios, with the exception of CLUSTERS, using a large data set from the National Medical Care Expenditure Survey. They found that the programs SESUDAAN and RATIOEST in the SUDAAN collection were the most efficient in terms of CPU time usage and easier to program than the others.

One major current trend in software development is the development of menu-driven packages on micro-computers. Variance estimation and specialized survey analysis software is no exception to this trend. A notable enhancement to the commonly used variance estimation programs since 1985 is the introduction of PC CARP (Schnell *et al.* 1986 and Schnell *et al.* 1988), available on IBM AT/XT or compatible micro-computers with a math co-processor. This package, like its predecessor SUPER CARP, uses Taylor linearization methods for variance estimation. A second variance estimation package is also available on micro-computers. The package listed as BELLHOUSE in Wolter (1985, p. 399) has been adapted for IBM micros with or without a co-processor by Rylett and Bellhouse (1988) under the program name TREES. This software uses tree structures to mimic the structure of stratified multistage sampling designs and applies tree traversal algorithms, in conjunction with general results on variance estimation in multi-stage sampling (see section 3.1), to the calculation of variance estimates.

A second trend in the computer implementation of survey variance estimation and survey analysis techniques is the integration of survey software with widely used statistical analysis systems. A leader in this trend from the early 1980's is the SUDAAN system, which is comprised of a series of several SAS procedures. Freeman *et al.* (1985) and Hidiroglou and Paton (1987) both used the PROC MATRIX procedure in SAS to obtain survey variance estimates, the former by balanced repeated replication and the latter by Taylor linearization. Mohadjer *et al.* (1986) report the development of a new SAS procedure WESVAR to obtain survey variance estimates by balanced repeated replication.

A variety of packages and computing techniques are available to carry out the analyses of survey data reviewed in Section 4. Among the available specialized packages, the most comprehensive appears to be the PC CARP. The original program, SUPER CARP, was designed to carry out regression analyses developed by Fuller (1975); the PC version retains this option. The current version now contains additional options for categorical data analysis, and inferences on cumulative distribution function and associated quantiles, following methods given by Francisco and Fuller (1986). For categorical data, there is an option for the analysis of two-way contingency tables, based on the Rao-Scott corrections to chi-squared test of independence. The program can also be manipulated to perform factor analyses of survey data.

There are four other specialized packages for the analysis of survey data; between them they cover topics in regression and categorical data analysis. The &REPER program in OSIRIS IV and the SURREGR procedure in SUDAAN both calculate standard errors of regression coefficients so that regression analyses can be carried out. The programs CPLX, developed by Fay (1982), and RSPLX, also by Fay, handle categorical data analyses of log-linear models for two and multi-way tables. The analysis in CPLX is carried out using jack-knifed chi-square statistics, while RSPLX applies second order Rao-Scott corrections to the usual test statistic.

The four programs for the regression analyses for complex survey data were evaluated by Cohen, Xanthopoulos and Jones (1988). The older version, SUPER CARP, was included in this analysis rather than PC CARP. Similar to the earlier study of Cohen, Burt and Jones (1986) on variance estimation, data from the National Medical Care Expenditure Survey were used. Once again, a program in the SUDAAN suite of programs, SURREGR, was the most

efficient in terms of CPU time usage and easier to program than the others. However, the efficiency of the SUDAAN programs might be balanced by the flexibility of the PC CARP program, depending upon the survey analysis required.

Significant enhancements to SUDAAN are provided in the new SUDAAN system under development (LaVange *et al.* 1989). Variance estimation and data analysis methods not available in SUDAAN are among the many modifications incorporated into the new SUDAAN System.

Running almost parallel to the emerging trend in the calculation of variance estimates, there is a move towards incorporating methods for the analysis of complex survey data into standard statistical packages and systems. Following on their variance estimation methods using SAS procedures, Hidioglou and Paton (1987) describe further SAS procedures to carry out log-linear analyses, with Rao-Scott corrections, of multi-way contingency tables. Likewise, Freeman (1988) notes that he used the SAS procedure PROC MATRIX for both variance estimation and for the analysis of variance of his survey data. Similarly, Mahodjer *et al.* (1986) describe two other new SAS procedures in addition to the variance estimation procedure WESVAR. These are the previously mentioned NASSREG and NASSLOG which carry out weighted least squares regression analyses and logistic regression analyses respectively. Both procedures depend on balanced repeated replication for variance estimation of the model parameters. An alternative approach to using SAS procedures is to use the matrix algebra language GAUSS (Platt 1986). Based on their own experience, Rao and Thomas (1988) favorably report on the use of this language for categorical data analysis in complex surveys.

6. CONCLUDING REMARKS

The early milestones in the development of efficient sampling designs and associated estimation techniques for population totals and means have firmly established sample survey theory and methods as a major discipline in statistics. Subsequent developments in theoretical foundations of sampling theory have provided useful insights into inferential aspects. In particular, the model-assisted approach and the conditional design-based approach appear to be promising since they attempt to fill the "gap" between the traditional approach and the model-dependent approach by retaining the desirable features of both approaches, but more research is needed in this area to handle complex sampling designs. Recent advances in variance estimation and confidence intervals for nonlinear statistics and the associated computer software, are also equally impressive. It is also gratifying that rapid progress has been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design, and the associated computer software.

We can expect to see important new developments in the next 10 years or so in the areas of variance estimation for nonlinear statistics (especially, nonsmooth functions), analysis of survey data (especially, multivariate analysis), and other topics not covered here (especially, sampling in time and small area estimation).

ACKNOWLEDGEMENTS

The authors would like to thank the editor for helpful comments. This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.
- BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis. In *The Analysis of Survey Data* (Eds. C.A. O'Muircheartaigh and C.D. Payne), Vol. 2, New York: Wiley, 175-192.
- BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.
- BELLHOUSE, D.R. (1988). A brief history of random sampling methods. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 1-14.
- BELLHOUSE, D.R., and RAO, J.N.K. (1986). On the efficiency of prediction estimators in two-stage sampling. *Journal of Statistical Planning and Inference*, 13, 269-281.
- BINDER, D.A. (1983). On the variance of the asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv.1, 6-62.
- BREWER, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.
- CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.
- CHAUDHURI, A. (1988). Optimality of sampling strategies. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 47-96.
- CHAUDHURI, A., and VOS, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam: North-Holland.
- COCHRAN, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COHEN, S.B., BURT, V.L., and JONES, G.K. (1986). Efficiencies in variance estimation for complex survey data. *American Statistician*, 40, 157-164.
- COHEN, S.B., XANTHOPOULIS, J.A., and JONES, G.K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data. *Journal of Official Statistics*, 4, 17-34.
- DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almqvist and Wiksell.
- DEMING, W.E. (1960). *Sample Design in Business Research*. New York: Wiley.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley-Interscience, 629-651.
- ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-224.
- FAY, R.E. (1982). Contingency tables for complex designs, CPLX. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 44-53.
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

- FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.
- FIELLER, E.C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428-440.
- FISHER, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd 5th edition 1934.
- FOLSOM, R.E. (1974). National assessment approach to sampling error estimation, sampling error monograph. National Assessment of Educational Progress, first draft.
- FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the distribution function with a complex survey. Technical Report, Iowa State University.
- FREEMAN, D.H. (1988). Sample survey analysis: analysis of variance and contingency tables. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 415-426.
- FREEMAN, D.H., LIVINGSTON, M., LEO, L., and LEAF, P. (1985). A comparison of indirect variance estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 313-316.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.
- FULLER, W.A. (1987). Estimators of the factor model for survey data. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 265-284.
- GHOSH, M. (1987). On admissibility and uniform admissibility in finite population sampling. In *Applied Probability, Stochastic Processes and Sampling Theory*, (Eds. I.B. MacNeil and G.J. Umphrey), Boston: D. Reidel Publishing Company, 197-213.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 17, 269-278.
- GODAMBE, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 28, 310-328.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society*, series B, 46, 270-272.
- GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.
- HÁJEK, J. (1981). *Sampling From a Finite Population*. New York: Marcel Dekker.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sampling Survey Methods and Theory*, Vol. 1. New York: Wiley.
- HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- HANSEN, M.H., DALENIUS, T., and TEPPING, B.J. (1985). The development of sample surveys of finite populations. In *A Celebration of Statistics: The ISI Centenary Volume* (Eds. A.C. Atkinson and S.E. Fienberg), New York: Springer Verlag, 327-354.
- HARTLEY, H.O., and RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R. (1980). *SUPERCARP-Sixth Edition*. Survey Section, Ames, Iowa.

- HIDIROGLOU, M.A., and PATON, D.J. (1987). Some experiences in computing estimates and their variances using data from complex survey designs. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 285-308.
- HOLT, D., and SMITH T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- HOLT, M.M. (1979). SURREG: standard errors of regression coefficients from sampling survey data. Research Triangle Institute, Research Triangle Park, North Carolina.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C.T. and FULLER, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.
- JONES, G.K. (1983). HESBRR (HES variance and crosstabulation program). Version 3, Internal NCHS Report, Hyattsville, Maryland.
- KIAER, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of Royal Statistical Society, series B*, 36, 1-37.
- KOCH, G.G., FREEMAN, D.H., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KOVAR, J., RAO, J.N.K., and WU, C.F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplement, 25-45.
- KOTT, P.S. (1987). Estimating the conditional variance of a design consistent regression estimator. Technical Report.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals Statistics*, 9, 1010-1019.
- LAVANGE, L.M., SHAH, B.V., BARNWELL, B.G., and KILLINGER, J.F. (1989). SUDAAN: A comprehensive package for survey data analysis. Technical Report, Research Triangle Institute.
- LEPKOWSKI, J.M. (1982). The use of OSIRIS IV to analyse complex sample survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 38-43.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- MADOW, W.G. (1978). Comments on papers by Basu and Royall and Cumberland. In *Survey Sampling and Measurement* (Ed. N.K. Namboodiri). New York: Academic Press, 315-322.
- MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-370.
- MCCARTHY, P.J. (1969). Pseudo-replication: half-samples. *International Statistical Review*, 37, 239-264.
- MOHADJER, L., MORGANSTEIN, D., CHU, A., and RHOADS, M. (1986). Estimation and analysis of survey data using SAS procedures WESVAR, NASSREG, and NASSLOG. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 258-263.
- MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

- NARAIN, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.
- NATHAN, G. (1988). Inference based on data from complex sample designs. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland.
- NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, series B, 42, 377-386.
- NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, series B, 12, 241-255.
- PFEFFERMAN, D., and HOLMES, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, series A, 148, 268-278.
- PFEFFERMAN, D., and SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.
- PLATT, W.G. (1986). GAUSS. *American Statistician*, 40, 164-169.
- RAJ, D. (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.
- RAO, J.N.K. (1971). Some thoughts on the foundations of survey sampling. *Journal of the Indian Society of Agricultural Statistics*, 23, 69-82.
- RAO, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators. *Journal of the Indian Statistical Association*, 17, 125-136.
- RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.
- RAO, J.N.K. (1987). Analysis of categorical data from sample surveys. In *New Perspectives in Theoretical and Applied Statistics* (Eds. M.L. Puri, J.P. Vilaplana and W. Wertz). New York: Wiley, 45-60.
- RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 427-447.
- RAO, J.N.K., and SINGH, M.P. (1973). On the choice of estimator in survey sampling. *Australian Journal of Statistics*, 15, 95-104.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, 231-241.
- RAO, J.N.K., and THOMAS D.R. (1988). The analysis of cross-classified categorical data from sample surveys. *Sociology Methodology*, 18, 213-269.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.

- ROYALL, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYALL, R.M., and HERSON, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 and 890-893.
- ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, series C, 37, 43-52.
- ROYALL, R.M., and PFEFFERMAN, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika*, 69, 401-410.
- ROYALL, R.M., and CUMBERLAND, W.G. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, series B, 50, 118-124.
- RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 4, 381-397.
- RYLETT, D.T., and BELLHOUSE, D.R. (1988). TREES: a computer program for complex surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 694-697.
- SÄRNDAL, C.E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRNDAL, C.E., and WRIGHT, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted regression technique for estimating the variance of the generalized regression estimator. *Biometrika*, 76, 527-537.
- SCHNELL, D., SULLIVAN, G., KENNEDY, W.J., and FULLER, W.A. (1986). PC CARP: Variance estimation for complex surveys. In *Computer Science and Statistics: Proceedings of the 17th Symposium of the Interface* (Ed. D.M. Allen). Amsterdam: North Holland, 125-129.
- SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.
- SCOTT, A.J., and SMITH, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.
- SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- SHAH, B.V. (1981a). SESUDAAN: Standard errors program for computing of standardized rates from sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.
- SHAH, B.V. (1981b)., RATIOEST: Standard errors program for computing ratio estimates for sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.
- SHAH, B.V. (1982). RTIFREQS: Program to compute weighted frequencies, percentages and their standard errors. Research Triangle Institute, Research Triangle Park, North Carolina.
- SKINNER, C.J., HOLMES, D.J., and SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- SMITH, T.M.F. (1984). Present position and potential developments: some personal views – sample surveys. *Journal of the Royal Statistical Society*, series A, 147, 208-221.
- THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

- TORTORA, R.D. (1980). The effect of disproportionate stratified design on principal component analysis used for variable elimination. *Proceedings of the Survey Research Section*, American Statistical Association, 746-750.
- VERMA, V., and PEARCE, M. (1977). Users manual for CLUSTERS: A sampling program for computation of sampling errors for clustered samples. Technical Report No. 568, World Fertility Survey, U.K.
- VINTER, S. (1980). Survey sampling errors with OSIRIS IV. *COMPSTAT 1980: Proceedings in Computational Statistics*, Vienna: Physica-Verlag, 72-80.
- WANG, J.C., and WU, C.F.J. (1988). An approach to the construction of asymmetrical orthogonal arrays. Technical Report, University of Waterloo.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures, *Journal of the American Statistical Association*, 47, 635-646.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.
- WU, C.F.J., HOLT, D., and HOLMES, D.J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, Third Edition, London: Griffin.

COMMENT

T.M. FRED SMITH¹

Sample surveys are one of the most important areas of the application of statistics. The paper by Professors Rao and Bellhouse is an excellent review of the theoretical development of sample surveys and I find it hard to be critical; but in the best traditions of the Royal Statistical Society I shall make the attempt in as constructive and a controversial manner as possible. In any review paper the choice of topics, especially relating to recent work, must be to some extent subjective. This affords a discussant an easy target; criticize the authors for their sins of omission. Also a review must be wide ranging and this allows discussants freedom to ride their own hobby horses over the range. I shall adopt both approaches and my objective in so doing is to identify some additional issues which I believe are important thus widening the review still further.

There is now general agreement about the milestones of our subject. These are associated with the names of Kiaer, Bowley, Neyman, Cochran, Hansen, Hurwitz, Madow, Mahalanobis, Horvitz and Thompson – an international collection dominated, latterly by contributions from the USA. Kiaer and Bowley's work was fundamental because they demonstrated that valid conclusions could be drawn from representative samples of quite small size drawn from large populations with arbitrary values. Representative samples were stratified samples with proportional allocation, and Bowley derived the appropriate theoretical results. Neyman and subsequent authors argued the case for random sampling and developed a comprehensive theory of randomisation inference applicable to most sampling schemes. Durbin (1953) completes the theory with his multi-stage sampling results. Despite the importance of these results sample surveys became a Cinderella subject on the fringes of mainstream statistics, and even today most university departments do not have a sampling statistician on their staff. Why is this?

One reason is that sample survey theory has developed mainly within social science and official government statistics, whereas most statisticians have a training within mathematics and physical science. Although all experimental scientists deal with samples very few seem to recognise this explicitly and those that do, such as geologists and biologists, have developed their own theory of sampling and estimation. In my view it is time to bring together sampling experts from all areas of scientific enquiry to share ideas and experiences and hopefully to establish a global theory of sample surveys.

A second reason is that sample surveys starts with a population which is a real fixed finite population of units. Samples are then drawn from this population according to specific rules. In most scientific enquiries the position is reversed; the population is not well defined and the scientist starts with a sample. One view of the role of the statistician, as enunciated, for example, R.A. Fisher, is to define the hypothetical population from which the sample data can be viewed as a random sample. This approach begs the question whether this hypothetical population has any scientific value. Arguably the sample survey approach of starting with the population has much to commend it.

A third reason is that since the finite population units can take arbitrary values the population cannot be summarized by a few parameters. Notions like sufficiency have little value in sample survey theory, and sample data are usually summarized by a mass of cross-tabulations. The estimation of a large number of cell proportions is the primary aim of sample surveys and the object of inference is usually descriptive rather than explanatory.

A final reason for the separation of sample surveys from mainstream statistics is that the randomisation theory of sample surveys is so complete. It is a closed theory which if accepted

¹ T.M.F. Smith, Department of Mathematics, The University, Southampton, SO9 5NH, U.K.

has few remaining problems to be solved. The chief concerns of randomisation researchers since Horvitz and Thompson (1952) provided the general theoretical framework have been the construction of π s sampling schemes with non-zero joint inclusion probabilities, the production of methods and programs for variance estimation and the construction of estimators which employ auxiliary information but can never be generally efficient because of Godambe's result. All of these problems are important, but they are not exciting, they lack the philosophical and mathematical depth to capture the imaginations of young mathematical statisticians.

These reasons are my explanation why sample surveys have been seen in the past as an activity on the fringe of mainstream statistics. The position is changing now and I detect a coming together of the branches of statistics. Much recent work in sample surveys has attempted to integrate surveys into mainstream statistics and many areas of statistics now recognise the importance of selection effects. Has the sample survey Cinderella been invited to the Statisticians's Ball?

In addition to his non-existence theorem Godambe has also shown that within the randomisation framework the likelihood is proportional to the probability of selection, $p(s | z)$, where z is the prior information on which the design was based, which for fixed s is a constant. Thus the likelihood is completely uninformative. In the same set-up Basu (1971) showed that the sufficient statistic is $\{(i, y_i) : i \in s\}$, namely the complete data tape including the labels. Although these results are also negative, highlighting the distinction between randomisation inference and other forms of inference, they did stimulate interest amongst a wider group of statistician and so had a positive value. My own interest in the theory of sample surveys was stimulated by Ericson (1969), in particular by the way he incorporated the uninformative likelihood into a positive framework via Bayes theorem and exchangeable priors. Ericson's use of exchangeability deserves consideration by all statisticians, not just Bayesians. Is it reasonable, is it even possible, to have a valid theory of predictive inference without some form of exchangeability? If there is no function of the unit values which is exchangeable how can you predict the unobserved values from the sample values? My opinion is that Ericson's work was a milestone in the development of sample survey theory.

The uninformative nature of the randomisation likelihood led some statisticians to question the role of randomisation. Godambe himself refers to "the problem of randomisation" and developed alternative theoretical approaches which required randomisation. Ericson also found a role for randomisation within his exchangeable set-up. He argued that if you employ your prior information, z , to form groups of units which are approximately exchangeable a priori then the use of simple random sampling will guarantee exchangeability. Royall (1970, 1973), however, made the mistake of advocating purposive sampling within his model-based framework. He touched a raw nerve and brought down upon his head the wrath of the randomisation establishment. I thought that Royall had asked some serious questions which deserved an answer and the strength of the reaction surprised me. Why did academic survey samplers and those from government agencies in North America feel so strongly about randomisation? Their colleagues in market research seemed happy with quota samples which could be viewed as a special case of balanced sampling. In Europe many official surveys are based on quota samples. What is so special about official statistics in North America?

I think the answer lies deep in the American political psyche. Thoughtful Americans are democratic in the true sense of that term. They believe in individual freedom and the right to information, they are also deeply suspicious of governments. They recognise the need within a democracy for reliable statistical information. To the official statisticians randomisation is the guarantee of the objective reliability of their data. It is a key source of their professional integrity and any attack on randomisation was seen as potentially dangerous however well

intentioned. I admire this position and it has helped to convince me that randomisation is one of the great contributions of statistics to science.

I have expressed myself with some feeling because I am so unhappy about the present position of official statistics in the U.K.. The tradition in the U.K. is not naturally democratic, we are still a monarchy, we respect authority rather than the individual. This tendency is being exploited and there is now a serious erosion of public confidence in the Government's use of statistics. It has been argued that official statistics in the U.K. are collected to aid the decisions of government, not to help parliament or to inform the electorate. Key series have been stopped, definitions have been changed, information is presented by ministers in ways which are patently false, yet no government statistician can complain publically because of the Official Secrets Act. There is a dangerous public cynicism about statistics and George Orwell's predictions in his novel 1984 may be closer to the truth than we realise. I apologise to the authors for this digression, but I said I would ride some hobby horses, and the issue of the integrity of official statistics is of great importance.

Before leaving randomisation theory I would like to make some comments about repeated surveys and rotation sampling. Again this is an area which the authors have excluded although they did note Patterson (1950) as a milestone paper. Randomisation theory has been developed within the framework of the one-off cross section survey. The extension to repeated surveys is non-trivial for it is difficult to retain the probability structure over time under rotation sampling when the population changes, Fellegi (1963). For the measurement of gross flows, or transition probabilities, the role of the randomisation inclusion probabilities is not clear. The beautiful simplicity of randomisation theory for one-off surveys is destroyed when they are repeated over time. But most important surveys are repeated surveys, especially in the government sector, so what are the implications?

As always the answer is that it depends. If the primary purpose is to produce descriptive statistics of the state of the system at each time period then the surveys can be considered as repetitions of a cross-section survey and each one can be analysed independently. Although composite estimators or time series estimators may be more efficient they should be viewed as secondary estimators rather than primary estimators. If I wanted to use repeated survey data within an econometric model I would prefer to input the cross-section estimates with their known correlation structure rather than complex composite estimates. On the other hand if I wanted the best estimate of the current value of, say, unemployment, for a particular purpose, not for public consumption, then I would use the most efficient procedure available. Similarly if I wanted to explain the change in value of some estimates over time then I would need to go beyond simple randomisation analysis. Thus the problems with randomisation inference for repeated surveys occur mainly for secondary analyses. However, there remains the important issue of which estimates should be reported to the public.

Section 2 of the paper is devoted to work on the theoretical foundations of inference from survey data carried out during the last 30-40 years. The authors have chosen to distinguish three approaches, design-based, model-dependent and model-assisted, the latter being an attempt to find a compromise solution between the other two. Personally I prefer to go for a GUT (Grand Universal Theory) approach integrating both design and models into one framework. The important influences on my thinking in this area, in addition to Ericson, have been Scott (1977) and Rubin (1976). In the GUT approach the survey variables, the sampling mechanism, and any other selection and measurement mechanisms are all introduced explicitly into an overall model. If Y is the $n \times p$ matrix of measured survey variables, z is the prior information, s denotes the sample, $s^* \subset s$ denotes the respondents, then the joint distribution of all these variables is

$$f(Y | z; \Theta)g(z; \phi)p(s | z)q(s^* | s, z, Y; \eta),$$

where the survey design, represented by $p(s | z)$, is of the so-called uninformative type such as random sampling. The design is uninformative because z is assumed known and includes all the usual information on stratification, clustering and measures of size. This general formulation forces statisticians to face up to all their assumptions. Non-response must be modelled explicitly. Measurement errors must be included in the structure of $f(Y | z; \Theta)g(z; \phi)$. The decision to use randomisation inference is then an explicit statement that given z the values of Y can be treated as unknown constants; they are arbitrary values about which we have no additional information. A modeller, on the other hand must specify the model to the level needed for inference, for example, by an exchangeable model. Both design-based and model-dependent approaches condition on the same prior information, z , and so both should employ similar, possibly identical, structures. In fact I would rarely expect the point estimators using the different approaches to differ very much in practice. The issue thus becomes that identified by the authors as the choice of a measure of uncertainty. Model-dependent procedures employ conditional variances, strict design-based procedures are unconditional. How to construct conditional design-based inferences is still an open question, but the approach of Robinson (1987) looks promising. The GUT model shows the design-based versus model-based controversy to be what it is, namely a relatively small philosophical dispute within the much bigger framework of total survey analysis.

The failure of both theoretical and practical statisticians to integrate sampling and non-sampling errors into measures of total survey error even after 50 years of intensive research must be noted as one of the failures of this important branch of statistics. But again things are changing and the mood now is no longer merely to report sampling errors and in addition to give vague warnings about the potential size of non-sampling errors but it is to attempt to measure total survey error recognising that some non-sampling biases can far exceed sampling errors.

Section 4 of the paper is devoted to the analysis of survey data, to the analytic rather than descriptive uses of surveys. Here the design-based, model-based dispute pales into insignificance. Analysts must face up to all the classical problems of model choice, estimation and testing, residual analysis and so on, which make up mainstream statistics. Cinderella is at last dancing with the Prince.

My final comments are again personal. If you look at the references at the end of the paper, and if you consider the additional areas which I have discussed, then you will see that Jon Rao has contributed important papers in every area. I think that it was particularly appropriate that he was invited to write this paper. I congratulate both authors on their fine paper.

ADDITIONAL REFERENCES

- DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society, Series B*, 15, 262-269.
- ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-233.
- FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā, C*, 39, 1-9.

A Historical Perspective on the Institutional Bases for Survey Research in the United States

STEPHEN E. FIENBERG and JUDITH M. TANUR¹

ABSTRACT

The basic theme of this paper is that the development of survey methods in the technical sense can only be well understood in the context of the development of the institutions through which survey-taking is done. Thus we consider here survey methods *in the large*, in order to better prepare the reader for consideration of more formal methodological developments in sampling theory in the mathematical statistics sense. After a brief introduction, we give a historical overview of the evolution of institutional and contextual factors in Europe and the United States, up through the early part of the twentieth century, concentrating on governmental activities. We then focus on the emergence of institutional bases for survey research in the United States, primarily in the 1930s and 1940s. In a separate section, we take special note of the role of the U.S. Bureau of the Census in the study of non-sampling errors that was initiated in the 1940s and 1950s. Then, we look at three areas of basic change in survey methodology since 1960.

KEY WORDS: Censuses; Cognitive aspects of survey design; Non-sampling errors; Probability sampling; Survey organizations.

1. INTRODUCTION

The development of survey methods in the technical sense can only be well understood in the context of the development of the institutions through which survey-taking is done. The purpose of this paper is to consider survey methods from this broader perspective in order to better prepare the reader for consideration of more formal methodological developments in sampling theory in the mathematical statistics sense that are described in numerous texts on sampling as well as in Rao and Bellhouse (1990). Although our viewpoint and organization is somewhat new, we have relied heavily on secondary sources which provide detailed expositions alternative to ours. Our paper focuses on the American experiences in the development of survey methodology, but it sketches some background of the much broader social science and institutional settings out of which survey methodology grew.

In the next section we present a very brief historical overview of the evolution of this institutional and contextual background, up through the early part of the twentieth century. We see two broad strands – social research and censuses. We begin with a short synopsis of the early history of European social research, turn to a brief overview of census-taking, especially in the context of the United States, and then take up the role of the International Statistical Congresses in the late nineteenth and early twentieth century in establishing the importance of sampling. Even following these congresses, the possible role of probability in sampling was not broadly understood. Further steps required an institutional base.

In section 3, we focus on the emergence of other U.S. institutional bases for survey research in the 1930s and 1940s. In particular, we note that a missing institutional ingredient was provided by the creation of the U.S. statistical agencies at the beginning of the twentieth century.

¹ Stephen E. Fienberg, College of Humanities and Social Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, U.S.A.; Judith M. Tanur, State University of New York, Stony Brook, U.S.A.

Then a number of factors, including the depression of the 1930s, the development of probability sampling methodology, and a U.S. federal statistical coordinating function came together to launch the modern era of survey methodology in the U.S. We also review market research and the universities as institutional bases. In section 4, we take special note of the role of the U.S. Bureau of the Census in the study of non-sampling errors that was initiated in the 1940s and 1950s. In section 5, we look at some of the basic changes in survey methodology since 1960, focusing on technological advances, the role of longitudinal surveys, and the recent movement to explore cognitive aspects of surveys.

2. A HISTORICAL OVERVIEW OF THE INSTITUTIONAL BACKGROUND FOR MODERN SURVEY METHODS

2.1 Institutional Bases in Early European Social Research

One set of roots of the U.S. tradition of survey methodology and data collection technology is in early European social research (*cf.* Lecuyer and Oberschall 1978, from whose work we have drawn).

In England that tradition can be traced to the seventeenth century. The research, dubbed political arithmetic, was based on administrative records (especially parish records) and personal observation. It was usually carried out by dedicated individuals, such as John Graunt who published his *Natural and Political Observations Made Upon the Bills of Mortality* in 1662. Until the beginning of the eighteenth century, the parish was the unit of local government and administration, so that it was sensible to use the clergy as informants for many inquiries. With the industrial revolution and the rise of cities, this convenient arrangement broke down, necessitating the institution of house-to-house surveys.

By the 1830s statistical societies were formed in England to investigate social problems. They organized committees, which in turn hired agents to go door-to-door to collect data. Although the statistical societies disbanded when the social problems seemed solved, similar procedures were revived towards the end of the nineteenth century when Booth (1889-1891) sent school attendance officers door-to-door to study London's poor.

In France, where the government was more highly centralized, early social research was carried out by the government. District administrators were used as informants to fill out questionnaires on the demographic and economic conditions of their districts. By the mid-eighteenth century what we might consider an early study of the effects of mass communication was carried out in France. Administrators were instructed to spread rumors of increases in taxes and of military conscription and to report on the reactions of the populace.

In the Napoleonic period following the revolution, the French government established a national office responsible for gathering survey-like data on population, social situation, agriculture, and industry and commerce (Bourguet 1988). While this effort was not fully successful, and while it fell short of census methods as we now understand them, it did set in place an institutional structure. During the nineteenth century, France continued the tradition of government responsibility for statistical functions through reporting of data by prefects and in its Bureau de Statistique. The Napoleonic effort also launched a social science data enterprise in France that explicitly rejected the ideas from the theory of probability as it was then known. The French interest in social statistics affected many scientists, such as Laplace and Quetelet (a Belgian who studied in France under Laplace), who in turn contributed in major ways to the art and science of census-taking, attempting to reintroduce ideas from probability, through the use of what we now know as ratio estimation (see Stigler 1986, Chapter 5). After

the revolution of 1830, the Académie des Sciences Morales et Politiques sponsored prize competitions that encouraged statisticians to undertake their own research.

In Germany, the origin of "statistics" (collection of data on the state) was in the universities as early as the end of the seventeenth century. By the early nineteenth century this work was split into three parts, with descriptive political science and historical/quantitative political economy remaining university-based but statisticians collecting data in census bureaus and other government agencies.

In 1872, the Verein for Socialpolitik was founded – part pressure group, part professional organization, part research organization. It drew up questionnaires to be answered by supposedly knowledgeable informants such as landowners, ministers, and notaries. Problems of informants' possibly inaccurate information, haphazardly grouped and imprecise questions, and low response rates dogged these efforts. By the early twentieth century, Levenstein (1912) published what was probably the first large scale attitude and opinion survey, for which he used a snowball sampling technique. At about the same time Max Weber attempted a survey of industrial workers, planning to get some information directly from respondents but finding that the majority did not care to cooperate.

2.2 Censuses: A Prelude to Survey-Taking

Another set of roots of survey methodology is intertwined with the history of methods for census-taking and thus we present a brief overview on censuses and census-taking infrastructures. Many others have observed that the origins of the modern census are found in biblical censuses described in the Old Testament (Madansky 1986) as well as in censuses carried out by the ancient Egyptians, Greeks, Japanese, Persians, and Romans (Taeuber 1978). The emphasis in the biblical accounts of censuses seemed to be on the results of the enumeration, rather than on how the counting was done, although in several instances we are told about the rapidity of the process. For most practical purposes we can skip from biblical times to the end of the eighteenth century and the initiation of census activities the United States of America, although there is some debate as whether Canada, Sweden, or the United States should be credited with originating the modern census (Willcox 1930).

In the United States, the first census was taken in 1790 (in 1990 the U.S. government will take its bicentennial census) by State officials who were then reimbursed by the Federal government. Then, in the next census of 1800, the enumerators were deputies or assistants to Federal marshals (Duncan and Shelton 1978). It was only with the 1880 census that the central Census Office gained control over field operations and secured the authority to appoint enumerators.

Prior to 1850 the U.S. decennial census considered the family as the unit of interest and reported few data on persons. The change to an individual-focus in census-taking was strongly influenced by the work Lemuel Shattuck, one of the of ASA's founders who had earlier conducted the Boston census of 1845 (Anderson 1988, pp. 36-37), as well as that of Quetelet, who helped to organize the 1846 Belgian census (Willcox 1930).

Progress on the methodology of census-taking continued, as every 10 years, a special operation was mounted to fulfill the constitutional obligation of an enumeration of the U.S. population; however, there was a clear lack of continuity from one census to the next (American Economic Association 1899). It was only after the first 12 censuses had been taken that the Bureau of the Census was created in 1902 as a permanent agency. Over this period there was a steady expansion of the number of censuses of other sorts and the broadening of topics covered in addition to simple enumeration.

2.3 International Statistical Congresses

The move from censuses to sample surveys was slow and laborious. Kruskal and Mosteller (1980) trace some of this movement, especially as it was reflected in the discussions regarding surveys that took place at the meetings of the International Statistical Institute (ISI), and our exposition here owes much to their work. The groundwork for these meetings was laid in the 1850s by Quetelet who helped to organize the first of a series of International Statistical Congresses in 1853. After nine such Congresses from 1853 to 1876, the ISI was founded in 1885. It is interesting to note that there is only one index entry for sample surveys in Stigler's (1986) history of statistics before 1900 – to 1830s work of Quetelet linked to a census method suggested by Laplace – and only two index entries in Porter's (1986) history – one to a 1900 paper by Karl Pearson and the other to the work of Kiaer and the ISI.

As early as the 1895 ISI meeting, Kiaer (1895-1896) argued for a “representative method” or “partial investigation”, in which the investigator would first choose districts, cities, *etc.*, and then units (individuals) within those primary choices. The choosing at each level was to be done purposively, with an eye to the inclusion of all types of units. That coverage tenet, together with the large sample sizes recommended at all levels of sampling, was what was judged to make the selection representative.

The idea of less than a complete enumeration was widely opposed, but Kiaer presented arguments for it (with some members agreeing and others disagreeing) at ISI meetings in 1897, 1901, and 1903. Towards the end of this period, the idea of probability sampling entered the discussion, but the topic of the representative method seems absent from the records of the ISI meetings until 1925. By then the record suggests that the representative method was taken for granted, and the discussions centered around how to accomplish representativeness and how to measure the precision of sample-based estimates (Bowley 1926; Jensen 1926). Notions of clustering and stratification were put forward, but purposive sampling was still the method of choice.

It was not until Gini and Galvani made a purposive choice of which returns of an Italian census to preserve and found that districts chosen to represent the country's average on seven variables were, in that sense, unrepresentative on other variables, that purposive sampling was definitively discredited (Gini 1928; Gini and Galvani 1929). Soon thereafter Neyman published his groundbreaking 1934 paper that demonstrated, among other thing, the virtues of probability sampling.

3. THE DEVELOPMENT OF INSTITUTIONAL BASES FOR SURVEY RESEARCH IN THE UNITED STATES

Survey research in the United States grew from a blending of the same three institutional bases that had been influential in Europe – private individuals acting as entrepreneurs in the private sector, universities, and the government. Early social research in this country (before World War I) seems to have followed the earlier British model, being carried out by social workers, public health workers, and reformers. An early university involvement was the hiring by the University of Pennsylvania in 1899 of W.E.B. DuBois to carry out his study of the Philadelphia Negro, conducted as a house-to-house survey. Starting in the 1930's, and especially in the period after World War II, the U.S. experienced a flowering of survey methodology in the three broad institutional bases: market research and polling, universities, and government. But before we describe that flowering we shall take a step backwards and note the establishment of the U.S. government statistical agencies.

Recently, Jean Converse (1987) has written an extremely scholarly and graceful study of the roots and emergence of survey research in the United States, with special focus on market research and polling and on universities. Our exposition on these bases closely follow hers. We have separated out the institutional bases both to reflect a social reality and to structure our exposition. But there is another social reality that we ask the reader to bear in mind; the membranes separating the institutions are permeable. They not only permit the flow of cross-fertilizing ideas and methods in all directions; they also permit a somewhat lesser flow of people, as individuals move from one sector to another over the course of their careers.

3.1 The Establishment of U.S. Statistical Agencies

The establishment of American statistical agencies effectively began in 1863, when the newly created Department of Agriculture released the first crop and livestock report to provide information on Union food supplies during the Civil War. This report was based on data from a purposive sample of 2,000 farmers in 22 states. This agricultural statistical reporting activity has existed in the Department of Agriculture on a continuing basis to the present day and is now known as the National Agricultural Statistical Service. By the late 1920s, correlational and regression methodology was well established in the work of agricultural statisticians (Duncan and Shelton 1978).

In 1884, Congress voted to establish a Bureau of Labor (later renamed the Bureau of Labor Statistics, BLS) to "collect information" on the earnings and the working conditions of "laboring men and women." Under the leadership of Carroll Wright, the first Commissioner, BLS expanded its statistical activities to cover such issues as depressions, strikes and lockouts, women's wages, marriage and divorce, and the domestic liquor trade (Norwood and Early 1984).

With the creation of the Bureau of the Census in 1902, there were three major U.S. agencies in place, each with a mandate to collect national data on a regular basis. During the first three decades of the twentieth century, the role of government statistical agencies expanded considerably and, at the time of the stock market crash of October 29, 1929, data on various facets of economic and social life were available. As late as 1932, however, there were few examples of probability sampling anywhere in the Federal Government (Duncan and Shelton 1978).

Difficult though it is to conceive in a period when we are used to receiving reliable readings on the unemployment rate monthly, there was no comparable survey data resource available in the 1920s and early 1930s. Except for selected monthly non-survey data gathered by BLS from most manufacturing industries and some nonmanufacturing industries, there were no regular national unemployment figures. In the 1920 census the question on unemployment was dropped because of statistical concerns regarding the accuracy of the resulting data. This question was restored to the 1930 census because of the wide-spread concerns regarding the employment situation. The extensive controversy that surrounded the 1930 unemployment data (Van Kleeck 1930) and those from the special January 1931 Unemployment Census was especially acrimonious (Anderson 1988), and played a role in the 1932 presidential election campaign.

3.2 The ASA-SSRC Committee and the Institutionalization of Probability Sampling: An Early Bridge

Thus, at the beginning of the Great Depression of the 1930s in the United States, the federal statistical agencies had difficulty responding to the demand for statistics to monitor the effects of the programs of President Franklin Roosevelt's New Deal. In 1933, Secretary of Labor Frances Perkins asked Stuart A. Rice, the ASA president, to set up an advisory committee on

the programs of BLS. This committee grew into the Committee on Government Statistics and Information Services (COGSIS), sponsored jointly by ASA and the Social Science Research Council (SSRC). Duncan and Shelton (1978) give a detailed account of the activities of COGSIS, and for our discussion here two outcomes are worthy of note.

First, in 1933, COGSIS recommended the creation of a Central Statistics Board (CSB) to help coordinate government statistical activities. With the groundwork laid for a coordinated federal statistical system, COGSIS and CSB proceeded, in early 1934, to arrange for an interagency agreement through which Census would collect basic data on production and labor for BLS.

Second, COGSIS helped to stimulate the use of probability sampling methods in various parts of the Federal government, and it encouraged research on sampling theory, to be done by employees of statistical agencies. For example, to establish a technical basis for unemployment estimates, COGSIS and CSB organized an experimental Trial Census of Unemployment as a Civil Works Administration project in three cities using probability sampling, carried out in late 1933 and early 1934. The positive results from this study and the interagency arrangement mentioned above led in 1940 to the first large-scale, ongoing sample survey on employment and unemployment using probability sampling methods. This survey later became the Current Population Survey.

Another somewhat indirect outcome of the COGSIS emphasis on probability sampling took place at the Department of Agriculture Graduate School where W. Edwards Deming organized a series of lectures in 1937 on sampling and other statistical methods by Jerzy Neyman (1938). These lectures had a profound impact on the further development of sampling theory across the government as well as in universities.

What we see happening in this period is the confluence of a number of factors that served to launch the use and development of sampling methods in the U.S. government statistical agencies. A key prerequisite was the existence of the agencies themselves. A second was the methodological advances in sampling theory as encapsulated in Neyman's landmark 1934 paper. What was required to bring these together was the Great Depression, a new administration hungry for quality data to assess the impact of its social programs, and the joint ASA-SSRC Committee on Government Statistics and Information Services.

3.3 Market Research and Polling

The institutional base of survey methodology in U.S. market research and polling traces its own pre-history to election straw votes collected by newspapers, dating back at least to the beginning of the nineteenth century. Often publicity and circulation boosting were more important than accuracy of prediction. Converse (1987) points out, however, a more serious journalistic base; election polls were taken and published by such reputable magazines as the *Literary Digest* (which had gained a reputation for accuracy before the 1936 fiasco). Then, as now, election forecasting was taken as the acid test of survey validity. A reputation for accuracy in "calling" elections was thought to spill over to a presumption of accuracy in other, less verifiable areas.

There was a parallel tradition in market research, dating back to just before the turn of the century, attempting to measure consumers' product preferences and the effectiveness of advertising. It was seen as only a short step from measuring the opinions of potential consumers about products to measuring the opinions of the general public about other objects, either material or conceptual. By the mid 1930s there were several well established market research firms. Many of them conducted election polls in 1936 and achieved much greater accuracy than did the *Literary Digest*. It was the principals of these firms (e.g., Archibald Crossley, George Gallup, and Elmo Roper) who put polling – election, public opinion, and consumer – on the map in the immediate pre-World War II period.

Data collection technology developed broadly in the market research and polling organizations in this era. Sampling was either by purposively selected groups or by quota. Samples were large, with the size enlarged sequentially until the law of large numbers caused the mean or percentage being estimated to stabilize. Some questionnaires were very informal with the interviewer instructed to bring certain topics into a conversation – what we might now call an unstructured interview. Others were more standardized, but shorter, actual forms. The progression seems to have been that as interviewers became more distanced from the primary investigators – in space, in education, in training, in identification with the research project, and perhaps in their very numerousness – the interview became more standardized.

The same kinds of validity issues that interest survey researchers today surfaced in the period. What should be the balance between open and closed questions? (Practice seems to have favored a combination; the device of the “opinion thermometer” to calibrate answers was first developed by the *Literary Digest* in 1925.) The pollsters tackled the problem of how to ask sensitive questions – about age, income, occupation, and home owning – by providing check lists, functioning much like contemporary visual aids. Experiments in question wording were carried out in the polling houses.

Market research in this early period, as now, of necessity put a premium on the timeliness of results. Then, as now, this tended to create some tension between academics and market researchers, with academics believing commercial workers to be corrupted by money and thus too far from basic science and commercial workers believing academics were overly concerned with the abstract. It is noteworthy, however, that one of the earliest homes of public opinion and market surveys was the Psychological Corporation, an organization of academic psychologists committed to plowing part of their profits back into the research process. The Psychological Corporation carried out its surveys from its Market Surveys Division, organized and run by Henry C. Link.

3.4 The Universities

But the universities were hardly totally above the polling movement. As early as 1911 the Harvard Graduate School of Business established a Bureau of Business Research to carry out consumer research. Such household names of social science as Paul Lazarsfeld, Hadley Cantril, and Rensis Likert moved to university affiliations and attached research institutes. Lazarsfeld came to the United States in 1933 determined to bring the techniques developed in market research to the basic scientific endeavor. He went on to form the Office of Radio Research, later to be called the Bureau of Applied Social Research, at Columbia University. His myriad contributions included the use of panels and a system of causal analysis.

Hadley Cantril was an academic who early on collaborated with Lazarsfeld on research on radio listening. When the two had a falling out, Cantril established the Office of Public Opinion Research at Princeton University. Here studies were carried out to improve data collection techniques. For example, in investigating the effects of question wording, Rugg and Cantril (1944) found that in 1940 – 41 over a six-week period, the percentage of Americans who favored “giving aid [to Great Britain] even at the risk of war” varied between 56% and 78%. At the same time, the percent in favor of “entering the war immediately” ranged from 8% to 22%.

Rensis Likert started out teaching at New York University and with a connection to the surveys of the Psychological Corporation. Moving to business, he carried out a survey of life insurance agents’ attitudes, comparing qualitative and quantitative (mostly questionnaire) methods. He then became Director of the Division of Program Surveys at the Department of Agriculture. There he worked to standardize questionnaires. When Likert left the Department of Agriculture after World War II, he brought his group to the University of Michigan to form the Survey Research Center.

4. FROM SAMPLING THEORY TO THE STUDY OF NON-SAMPLING ERROR

As we have seen above, the introduction of probability sampling into government surveys in the mid-1930s came at the time of rapid development in many areas of statistics, and the development of a foundation for experimentation and inference more broadly under the leadership of such statisticians as R.A. Fisher, Walter Shewart, Jerzy Neyman, and Egon Pearson. Among those who worked on the probability-sampling-based trial Census of Unemployment at the Bureau of the Census were Calvert Dedrick, Morris Hansen, Samuel Stouffer, and Frederick Stephan (Anderson 1988; Duncan and Shelton 1978). Hansen was then assigned with a few others to explore the field of sampling for other possible uses at the Bureau, and went on to work on the 1937 sample Unemployment Census. After working on the sample component of the 1940 decennial census (under the direction of Deming), Hansen worked with others (e.g., Jerome Cornfield, Lester Frankel, William Hurwitz and J. Steven Stock) to redesign the unemployment survey based on new ideas on multi-stage probability samples and cluster sampling (Hansen and Hurwitz 1942, 1943). They expanded and applied their approach in various Bureau surveys, often in collaboration and interaction with others, and this effort culminated in 1953 with the publication of a two-volume compendium of theory and methodology (Hansen, Hurwitz and Madow 1953). The recent interview with Hansen (Olkin 1987) and the Duncan and Shelton (1978) volume provide interesting and detailed descriptions of the developments during this period.

Virtually independent and often complementary contributions to sampling theory came via the statistical sampling work in agriculture by P.C. Mahalanobis and students in India and by Frank Yates and William Cochran in England. Cochran's 1939 paper is especially notable because of its use of the analysis of variance in sampling settings and the introduction of superpopulation and modeling approaches to the analysis of survey data (see Fienberg and Tanur 1987, 1988 for related discussion on the design and analysis linkages between sampling and experimentation). In the 1940s, as results from these two separate schools appeared in various statistical journals, we see some convergence of ideas and results.

The 1940s saw a rapid spread of probability sampling methods to other government agencies. It was only after the fiasco of the 1948 presidential pre-election poll predictions (Mosteller *et al.* 1949) that market research firms and others shifted towards probability sampling. Even today many organizations use a version of probability sampling with quotas (Sudman 1987).

Amidst the flurry of activity on the theory and practice of probability sampling during the 1940s, attention was also being focused on issues of nonresponse and other forms of non-sampling error. In a review of work on errors in surveys, Deming (1944) listed 13 factors affecting the ultimate usefulness of surveys (note that most of these are nonsampling errors):

1. variability in response;
2. differences between different kinds and degrees of canvass;
3. bias and variation arising from the interviewer;
4. bias of the auspices;
5. imperfections in the design of the questionnaire and tabulation plans;
6. changes that take place in the universe before tabulations are available;
7. bias arising from nonresponse (including omissions);
8. bias arising from late reports;
9. bias arising from an unrepresentative selection of date for the survey, or of the period covered;

10. bias arising from an unrepresentative selection of respondents;
11. sampling errors and biases;
12. processing errors (coding, editing, calculating, tabulating, tallying, *etc.*);
13. errors in interpretation.

Most of the errors described in this list either had been or would become the focus of research by statisticians at the Bureau of the Census.

A milestone in this effort to understand and model non-response errors was the development of an integrated model for sampling and non-sampling error in censuses and surveys, in connection with planning for and evaluation of the 1950 census (Hansen, Hurwitz, Marks and Mauldin 1951). This analysis-of-variance-like model, or variants of it, has served as the basis of much of the work on non-sampling error over the past 35 years, both inside and outside the Bureau of the Census. An excellent qualitative analysis of the error structure of the Current Population Survey is given in Brooks and Bailer (1978), and reviews of the non-sampling error literature are given by Mosteller (1978) and Fienberg and Tanur (1983). Finally, we note that Groves' (1989) recent book gives an updated approach to a variant of this census model, making a careful distinction between random and fixed components that arise from the various sources of error.

The paper by Bailer (1990) in this issue contains a detailed discussion on non-sampling error from the perspective of the Bureau of the Census.

5. CHANGING DIMENSIONS OF SURVEY METHODOLOGY AFTER 1960

The decades of the 1960s and 1970s saw polls and surveys becoming an all-pervasive fact of American life, beginning with the hard-fought presidential election of 1960 in which both candidates (Kennedy and Nixon) commissioned and relied on private polls of the electorate. Here we focus on three major areas of innovation during recent decades. We refer the reader to other presentations for such important topics as imputation for incomplete data and the ever-present controversies surrounding inferences from survey data (*e.g.*, see Fienberg and Tanur 1983, 1986).

5.1 Mode of Interviewing: The Role of Telephones and Computers in Surveys

The development and diffusion of technology, especially telephones and computers, strongly influenced survey practice in these decades. U.S. telephone coverage, which was estimated to have been only 35% in 1936 and hence contributed to the *Literary Digest's* problem (Massey 1988), reached 75% by 1960 and 88% in 1970 on its way to around 93% in 1986 (Thornberry and Massey 1988). Thus telephone surveys, often based on random digit dialing (RDD) techniques, became increasingly prevalent and accurate. The movement began among commercial survey researchers, with governmental and academics lagging behind because of their concerns over differential coverage by such variables as income and race (Trewin and Lee 1988) and accompanying fears of lack of "representativeness". Indeed, most government uses of telephone interviewing remain as follow-ups of initial in-person contacts (as in the Current Population Survey which has been using telephone interviewing for households in later months of the survey since 1954). Only recently has there been a marked shift towards the use of RDD for government surveys. Groves and Kahn (1979) provide a review of work on telephone interviewing and, by and large, they document the comparability of survey results through comparisons of data gathered by personal interviews and by telephone.

The advent and proliferation of the computer meant that the tasks of analyses of survey responses could be carried out much more rapidly and broadly than ever before. This led to an increase in the number of surveys carried out under all institutional auspices. In retrospect it seems only natural that computer technology should be combined with telephone technology to produce systems of computer assisted telephone interviewing (CATI). These systems provide automated questionnaires that carry out skip patterns and display the appropriate question on a monitor screen, schedule (and often actually place) calls and callbacks, carry out randomizations, and automate data entry, in addition to other functions. CATI systems were developed by U.S. market research organizations in the early 1970s in part to keep track of respondent characteristics and thus ensure that quotas are precisely and efficiently met (Nicholls 1988). Chilton Research was one of the commercial CATI pioneers, using a CATI system for surveys intended to determine the level of customer satisfaction with services provided by the telephone companies (Nicholls and Groves 1986). Largely independently, university survey organizations began to develop their own CATI systems in the mid-1970s, and introduced them to the larger statistical community with an emphasis on their usefulness for documentation, standardization, and interviewer flexibility. While government agencies exhibited early interest in CATI, they have only recently begun to actually employ systems, sometimes on an experimental basis and often in tandem with other data collection methodologies, as in panel designs where the first interview is carried out in person. At this writing we see the beginnings of a movement to the use of computer-assisted personal interviewing (CAPI), a development made possible by the technological advances that produced truly portable laptop computers.

5.2 Longitudinal Surveys

While panel surveys were conducted in connection with the 1924 and 1940 U.S. presidential election campaigns (Rice 1928; Lazarsfeld *et al.* 1944), interest in over-time survey data did not really become fashionable in social research until the 1960s. This is all the more surprising when we realize that the Current Population Survey has traditionally had a rotating-panel structure and, since 1953, many respondents are interviewed as many as 8 times over a 16 month period. This rotating-panel structure was originally intended to produce estimates of change in aggregate quantities that had smaller variances than those from repeated cross-sections but, in principle, the CPS could have been analyzed in panel form on a regular basis. The fact that the CPS is a survey of sample addresses and not individuals or households is a major obstacle to the use of it as a panel survey (see related comments on the National Crime Survey in Fienberg 1978), but this has not prevented the elaborate use of the CPS to study gross flows in individual employment status (*e.g.*, see Abowd and Zellner 1985 and Stasny 1988).

Not all survey attempts to measure change need be based on longitudinal data; often repeated cross-sections can do at least as well if not better in measuring aggregate change. By the 1970s the Gallup Poll and others had developed a tradition of asking the same questions repeatedly and reporting the results in newspapers. These established time series became incorporated into the burgeoning Social Indicators movement. In 1972 the National Opinion Research Center first fielded the General Social Survey (GSS), funded by the National Science Foundation. GSS was designed by a broadly based group of academics to provide periodic readings on social indicators and to provide an original data set for use by students and academics doing modestly funded research. For purposes of continuity, the designers incorporated into GSS many questions first developed by Gallup and other commercial pollsters, yielding a fruitful cross-institutional collaboration (*e.g.*, see Smith 1975).

The basic idea behind the conduct of longitudinal surveys of panels, however, is to measure changes over time, not by comparing the changes in aggregate quantities, but by focusing on individual change. Such surveys typically focus on changes in status, the duration of activities, and events occurring over time. The rise of interest in longitudinal panel surveys occurred primarily outside the government, and early examples are the Panel Study of Income Dynamics, conducted by the Institute for Social Research at the University of Michigan annually since 1968; the National Longitudinal Surveys of Labor Market Experience, sponsored by the Center for Human Resources Research at Ohio State University beginning in 1966, and currently funded by BLS; and the Longitudinal Retirement History Survey, sponsored by the Social Security Administration from 1969 to 1979. The 1970s saw expanded use of longitudinal panel surveys, especially under government auspices (*e.g.*, see Boruch and Pearson 1988), but the basic survey methodology used often resembled that for traditional cross-sectional surveys. Only in the late 1970s did researchers begin to question the conventional wisdom about longitudinal survey design and analysis and to explore such fundamental issues as the definition of a longitudinal family (for a discussion, see Fienberg and Tanur 1986).

In the 1980s, interest in longitudinal panel surveys expanded and considerable attention was focused on aspects of non-sampling error such as attrition and on issues of data management and analysis. Kalton *et al.* (1989) includes a number of papers on these topics.

5.3 Cognitive Aspects of Surveys

As a result of systematic efforts to improve survey methodology over the past forty years, survey researchers have evolved a highly developed art of questionnaire design and interview procedures to reduce nonsampling errors, such as those described in Deming's list above (*e.g.*, see Payne 1951), and they have carried out many scientific studies to test aspects of that art (*e.g.*, see Sudman and Bradburn 1974, Bradburn and Sudman 1979, and Schuman and Presser 1981). Until recently, however, research on understanding the survey interview situation has been relatively unsystematic. The recent change came, in part, through the recognition that other fields, in particular cognitive psychology, had insights that would assist survey researchers in examining the interview process.

Among non-sampling errors are those occasioned by the cognitive processes that respondents and interviewers are required to exercise in the survey interview situation. Respondents must often recall events and make judgments or estimates, and must always face issues of comprehension of the questions asked – their meaning to respondents as well as their meaning to interviewers. Survey researchers are now beginning to draw on the concepts of cognitive psychology and the expertise of cognitive psychologists to investigate more systematically these issues of non-sampling error. We note especially that the exploration of meaning is not new to the enterprise of survey research. Indeed, Cantril (1944) devotes two chapters to reporting the results of experiments on the meaning and wording of questions. These experiments used many of the same probing and paraphrasing techniques used in today's cognitive laboratory.

This explicit movement to study cognitive aspects of surveys originated in a 1981 conference sponsored by the Bureau of Social Science Research and the Bureau of Justice Statistics that brought together cognitive psychologists and survey researchers to concentrate on the National Crime Survey. A more intensive 1983 conference, sponsored by the Committee on National Statistics (CNSTAT) of the National Research Council, concentrated on the National Health Interview Survey (Jabine *et al.* 1984). From the beginning the movement was, by design, a partnership between people from academia, from research institutes and other academic institutions, and from the government.

A direct outgrowth of the CNSTAT conference was the establishment of a Questionnaire Design Research Laboratory at the U.S. National Center for Health Statistics under the leadership of Monroe Sirken to do pretesting (in parallel with full scale field testing) of major government surveys. It employs government personnel, brings in visiting scholars, and contracts with academics and people in research institutes to carry out its mission. This has been followed by the establishment of similar laboratories at the Bureau of Labor Statistics and the Bureau of the Census. Another outgrowth is the establishment of the Social Science Research Council's Committee on Cognition and Survey Research, which is, itself, both cross disciplinary and cross institutional. The Committee has fostered research in such directions as the interactive process of the survey interview, the uses and pitfalls of retrospective memory, and issues in measuring pain in a survey context. Examples of other outgrowths of this movement are (a) an investigation by the OECD's Working Party on Labor Statistics of cognitive aspects of labor surveys, addressing such issues as the meaning of "looking for work" – a knotty conceptual problem within a culture, and even more problematic across cultures (Schwarz 1987), (b) work at combining the cognitive perspective with statistical work on the embedding of experiments within surveys (Fienberg and Tanur 1989), (c) international conferences on work at the interface of cognition and survey methods (*e.g.*, see Hippler, Schwarz and Sudman 1987).

At the same time that methodological techniques of the cognitive laboratory are being used to shape questionnaire design, findings from the cognitive psychology laboratory are being taken into the field in order to test their generalizability and thus enrich the academic field of cognitive psychology, as well as to ascertain their usefulness for the survey enterprise. Here is yet another instance of interaction between the academic world and the government. For example, a laboratory finding is that people recall visits to health care providers more easily and accurately if they begin with the earliest first (Fathi, Schooler and Loftus 1984). A recent investigation explores whether this advantage holds in the field situation of the pre-test of the NHIS (White and Berk 1987).

The movement to integrate methods from the cognitive sciences into the design of sample surveys is important for several reasons. First, it has brought a renewed scientific base to the problems of questionnaire design. Second, it has opened up the survey domain to the study of selected cognitive phenomena. But most important, it had brought new vigor to the survey enterprise and raised anew issues about the structure and format of the survey interview, going far beyond questionnaire design, that many statisticians thought were resolved in the 1940s and 1950s.

6. COMMENTS

Traditional reviews of the history of survey methods have focused on the role of probability sampling and its refinements, and occasionally on the study of non-sampling errors. Here we have attempted to set this methodological history in the context of the tradition of social science research that evolved over the nineteenth and early twentieth centuries and the institutions, in and outside of government, that facilitated and occasionally directly spawned the methodological developments. This perspective should help remind readers that factors other than the advance of statistical theory have helped to shape the survey domain as we know it today. It should also help them follow the evolution of survey theory and practice as it continues to be shaped by institutional change.

There is an additional facet of institutional shaping of the survey enterprise that we have not addressed heretofore. We wrote above about the permeability of the membranes separating the three sectors: government, market research (the private domain), and the universities and

other academic institutions. We believe that these membranes are becoming even more permeable with the increased presence of a fourth kind of institution, which we shall refer to as a "bridge". We saw earlier how the ASA-SSRC Committee on Government Statistics and Information Services, a bridge between academia and government, prepared the ground for federal statistical coordination. ASA and SSRC continue to provide bridging functions, but other such institutions also exist.

Some vivid examples of other bridges come to mind. For over 40 years the American Association for Public Opinion Research has been bringing together survey practitioners from all sectors in local chapters and in national conferences at which new findings are disseminated and issues of common concern are discussed. The National Science Foundation program on Measurement Methods and Data Improvement (MMDI), under the direction of Murray Aborn, has explicitly seen as part of its mandate the fostering of government/academic collaboration. The mission has been implemented, for example, through the funding of research by academics that both uses and improves government databases (the 1983 seminar on cognitive aspects of survey methodology was sponsored by MMDI) and the funding of an ASA-sponsored fellowship program. That fellowship program places academic researchers for a semester or a year in government statistical agencies to carry out their own research, bring new ideas to the agency, and return to their academic bases with new knowledge and contacts in the federal agencies and new awareness of government data bases and statistical concerns. The National Research Council, an arm of the National Academy of Sciences, maintains a Committee on National Statistics that brings statisticians from academia and the private sector together to interact with representatives of the government agencies. Here, in formal panel studies and informal interaction, individuals come to know one another and common problems are tackled.

While these and other bridges will surely not totally erase the boundaries between the sectors, we see their existence as a positive force for progress in the development of survey methodology. Developments in one sector move more quickly to others across these bridges, but perhaps more important, the bridges facilitate a process whereby problems faced by any sector become legitimate research questions in all sectors.

ACKNOWLEDGEMENTS

The preparation of this paper was supported in part by the National Science Foundation under Grant No. SES-8701606 to Carnegie Mellon University and Grant No. SES-8701816 to the State University of New York at Stony Brook. An earlier version appeared under the title "Some History of Survey Methods and Data Collection Technology," in the *Sesqui-centennial Invited Paper Sessions* volume of the 1989 Proceedings of the American Statistical Association, 393-405.

REFERENCES

- ABOWD, J.M., and ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- ANDERSON, M.J. (1988). *The American Census. A Social History*. New Haven: Yale University Press.
- AMERICAN ECONOMIC ASSOCIATION (1899). The Federal Census. Report of the committee on the Twelfth Census. *Publications of the American Economic Association*, New Series, No. 2, 1-7.

- BAILAR, B.A. (1990). Contributions to statistical methodology from the federal government. *Survey Methodology*, 16, 51-61.
- BOOTH, C. *et al.* (1889-1891) 1902-1903. *Life and Labours of the People in London*. London: Macmillan.
- BORUCH, R.F., and PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.
- BOURGUET, M.-N. (1988). Décrire, Compter, Calculer: The debate over statistics during the Napoleonic Period. In *The Probabilistic Revolution. Volume 1, Ideas in History* (Eds. L. Kruger, L.J. Daston and M. Heidelberger). Cambridge: MIT Press.
- BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, 6-62.
- BRADBURN, N.M., SUDMAN, S., and Associates (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- BROOKS, C.A., and BAILAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Working Paper 3, Office of Federal Statistical Policy and Standards. Washington: U.S. Department of Commerce.
- CANTRIL, H. (1944) (1947). *Gauging Public Opinion*, Princeton: Princeton University Press.
- COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- CONVERSE, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.
- DEMING, W.E. (1944). On errors in surveys. *American Sociological Review*, 19, 359-369.
- DUBOIS, W.E.B. (1899) (1973). *The Philadelphia Negro: A Social Study; Together With a Special Report on Domestic Service by Isabel Eaton*. Millwood, N.Y.: Kraus Reprint.
- DUNCAN, J.W., and SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. U.S. Department of Commerce. Washington: U.S. Government Printing Office.
- FATHI, D., SCHOOLER, J., and LOFTUS, E. (1984). Moving survey problems into the cognitive psychology laboratory. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 19-21.
- FIENBERG, S.E. (1978). Victimization and the National Crime Survey: Problems of design and analysis. In *Survey Sampling and Measurement* (Ed. K. Namboodiri). New York: Academic.
- FIENBERG, S.E., and TANUR, J.M. (1983). Large scale social surveys: Perspectives, problems, and prospects. *Behavioral Science*, 28, 135-153.
- FIENBERG, S.E., and TANUR, J.M. (1986). The design and analysis of longitudinal surveys: Controversies and issues of cost and continuity. In *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits* (Eds. R.W. Pearson and R.F. Boruch). New York: Springer-Verlag.
- FIENBERG, S.E., and TANUR, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Canadian Journal of Statistics*, 55, 75-96.
- FIENBERG, S.E., and TANUR, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *Canadian Journal of Statistics*, 16, 135-151.
- FIENBERG, S.E., and TANUR, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 242, 1017-1022.
- GINI, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population Italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23, 198-215.
- GINI, C., and GALVANI, L. (1929). Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione (1 dicembre 1921) (In Italian). *Annali di Statistica*, Series 6, 4, 1-107.

- GRAUNT, J. (1662) (1939). *Natural and Political Observations Made Upon the Bills of Mortality* (Edited and with an introduction by Walter F. Willcox). Baltimore: Johns Hopkins Press.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L. II, and WAKSBERG, J., eds. (1988). *Telephone Survey Methods*. New York: Wiley.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.
- GROVES, R.M., and KAHN, R.L. (1979). *Surveys by Telephone*. New York: Academic Press.
- HANSEN, M.H., and HURWITZ, W.N. (1942). Relative efficiencies of various sampling units in population inquiries. *Journal of the American Statistical Association*, 37, 89-94.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, New York: Wiley.
- HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., and MAULDIN, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- HIPPLER, H.-J., SCHWARZ, N., and SUDMAN, S., eds. (1987). *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.
- JABINE, T.B., STRAF, M., TANUR, J.M., and TOURANGEAU, R., eds. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington: National Academy Press.
- JENSEN, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, 359-380.
- KALTON, G., KASPRZYK, D., and DUNCAN, G.J., eds. (1989). *Panel Surveys*. New York: Wiley.
- KIAER, A.N. (1895-1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute*, 9, 176-183.
- KRUSKAL, W.H., and MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48, 169-195.
- LAZARSFELD, P.F., BERELSON, B., and GAUDET, H. (1944). *The People's Choice: How the Voter Makes up his Mind in a Presidential Campaign*. New York: Columbia University Press.
- LECUYER, B., and OBERSCHALL, A. (1978). Social research, the early history of. In *International Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.
- LEVENSTEIN, A. (1912). *Die Arbeitfrage mit besonderer Berücksichtigung der sozialpsychologischen Seite des modernen Grossbetriebes und der psychophysischen Einwirkungen auf die Arbeiter*. (In German) Munich: Reinhardt.
- MADANSKY, A. (1986). On biblical censuses. *Journal of Official Statistics*, 2, 561-569.
- MASSEY, J.T. (1988). An overview of telephone coverage. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.
- MOSTELLER, F. (1978). Errors: 1. Nonsampling errors. In *International Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.
- MOSTELLER, F., HYMAN, H., MCCARTHY, P.J., MARKS, E.S., and TRUMAN, D.B. (1949). *The Pre-election Polls of 1948*. Bulletin 60. New York: Social Science Research Council.
- NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- NEYMAN, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, DC: Graduate School, U.S. Department of Agriculture.
- NICHOLLS, W. L., II (1988). Computer-assisted telephone interviewing: A general introduction. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.

- NICHOLLS, W.L., II, and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part 1 – Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.
- NORWOOD, J.L., and EARLY, J.F. (1984). A century of methodological progress at the U.S. Bureau of Labor Statistics. *Journal of the American Statistical Association*, 79, 748-761.
- OLKIN, I. (1987). A conversation with Morris Hansen. *Statistical Science*, 2, 162-179.
- PAYNE, S. L. (1951). *The Art of Asking Questions*. Princeton: Princeton University Press.
- PORTER, T.M. (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton: Princeton University Press.
- RAO, J.N.K., and BELLHOUSE, D.R. (1990). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *Survey Methodology*, 16, 3-29.
- RICE, S. (1928). *Quantitative Methods in Politics*. New York: Knopf.
- RUGG, D., and CANTRIL, H. (1944) (1947). The wording of questions. In *Gauging Public Opinion* (Ed. H. Cantril). Princeton: Princeton University Press.
- SCHWARZ, N. (1987). Cognitive aspects of labor surveys in a multinational context. Paper prepared for the Working Party on Labor Statistics, OECD. Paris, April 1987.
- SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic.
- SMITH, T.W. (1975). Social change and the General Social Survey: An annotated bibliography. *Social Indicators Research*, 2, 9-38.
- STASNY, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating gross labor force flows. *Journal of Business and Economic Statistics*, 6, 207-219.
- STIGLER, S.M. (1986). *The History of Statistics. The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.
- SUDMAN, S. (1987). *Reducing the Costs of Surveys*. Chicago: Aldine.
- SUDMAN, S., and BRADBURN, N.M. (1974). *Response Errors in Surveys: A Review and Synthesis*. Chicago: Aldine.
- TAEUBER, C. (1978). Census. In *International Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.
- THORNBERRY, O.T. Jr., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.
- TREWIN, D., and LEE, G. (1988). International comparisons of telephone coverage. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.
- VAN KLEECK, M. (1930). The Federal Unemployment Census of 1930. *Proceedings of the American Statistical Association*, 189-200.
- WHITE, A.A., and BERK, M.L. (1987). Recall strategies in personal interviewing: moving results from the laboratory to the field. *Proceedings of the Social Statistics Section, American Statistical Association*, 66-71.
- WILLCOX, W.F. (1930). Census. In *Encyclopaedia of Social Sciences* (Eds. E.R.A. Seligman and A. Johnson). New York: Macmillan.

COMMENT

ROBERT M. GROVES¹

The writing of histories of the development and use of survey methods signals a certain maturation of the field. Currently, we are seeing the fiftieth anniversary of several important survey innovations – Neyman's breakthrough papers in stratification, the start of the U.S. Current Population Survey, and the greater visibility of election polling. With our attention called to such developments it is natural to review the intervening years, seeking to find some theme for events affecting the field. Professors Fienberg and Tanur have completed such an exercise in their paper.

My comments will review key parts of the work, offering comments as I proceed, and then note some errors of nonobservation, misplaced emphases, and other minor quibbles.

Fienberg and Tanur express the purpose of their paper in two ways "to note that technical developments in surveys can be understood only in the context of institutions within which they occur" (p. 31) or at another point to note that "factors other than the advance of statistical theory have shaped the survey domain" (p. 42). Consistent with this they note:

1. the role of ruling, governing institutions which perceive a need for information on the population's welfare or its reaction to taxation;
2. later, the role of academics in the social sciences in framing central statistical and measurement issues in surveys;
3. the role of mass media use of surveys for election and current events monitoring; and
4. still later, the use of surveys by commercial entities in the market economy.

They document the resolution of controversies in the government sector about use of probability sampling.

Along the way we learn some interesting facts – for example, that for 12 U.S. censuses (120 years) there was no permanent organization for the Census Bureau; that the Department of Agriculture data collection began with need for information about food supplies in the Civil war; that another boost for surveys occurred in the New Deal's creation for government programs. There seems to be a recurring theme here that governments emphasizing services for the welfare of the populace demand more information about their societies than do those pursuing other goals. In addition, we see that governments most sensitive to public opinion demand more measures of that opinion (that reminds this reader of Gallup's early metaphor of the survey as a voting analog).

The focus in the paper on the role that institutions played in the development is convincing only for parts of the review. For example, the institutional focus is appealing in describing Lazarsfeld's evangelical efforts to bring commercial survey and academic inquiry together. The role of the Bureau of Applied Social Research at Columbia University in his partial success at that is enlightening. So too the move of Likert and others from a government agency home (U.S. Department of Agriculture) to academia in order to spread the method to new domains is largely a story of groups of people and organizations which make them effective.

However, the identification of organizations or institutions as the focus can be misunderstood as the stimulus to developments. Nothing I read in the paper changes my opinion that the survey field at its origins attracted broad, creative thinkers. Many were intelligent and charismatic; they led by ideas and mobilized others to work diligently at the definition

¹ Robert M. Groves, The University of Michigan and U.S. Bureau of the Census.

of the new field. Institutions permitted this to happen. They didn't produce the developments. They were homes for the best and brightest.

Within the focus on the institutional, I wished the emphasis of the paper might have been placed more on two related points:

1. Different tasks were more easily accomplished in the different domains. For example, government agencies were by their nature restricted to questions of monitoring social welfare, the commercial, to newsworthy or dollar worthy interests, and the academic to longer term, more basic social issues. Those involved in early developments shaped their agenda to the goals of the organization.
2. Stories of the early days of survey research, as told by those who lived them are filled with the excitement of a new field. I missed in the paper sufficient acknowledgement that the young researchers involved in the work shared an evangelical mission – spreading the gospel of probability sampling, inventing new methods of interviewing because nothing existed. The institutional focus misses the human drama of those days.

Fienberg and Tanur also note “the membranes separating the institutions are extremely permeable”. That is, researchers move back and forth between the institutions, contributing to each of them, and transferring knowledge as they move. The evidence the authors cite is the experience of Lazarsfeld addressing basic design issues while conducting radio audience research within an academic setting and of Likert moving from the insurance industry to the Department of Agriculture to the University of Michigan. These moves seem the exception rather than the rule. I have not conducted the appropriate careerline research to demonstrate this, but my impression is that the fences between the sectors have been and remain high and painful to transgress. Further, movement among academic government, commercial is asymmetric. Rarely is there movement from the commercial or government sector to the academic sector (current demands on publication history prevent this). The government-commercial interchange is larger.

The result of this insularity is the development of techniques not shared across the different sectors (edit and imputations schemes, nonresponse reduction techniques). The three sectors to some extent have developed their own language to describe their work (*e.g.*, “stem and banners”, “tabs” versus “contingency tables”).

The membrane metaphor also fails to observe the large differences in the centrality of surveys to organizations in the three sectors. Academic survey research is not central to any university in the world. It was not central early in the history of the method (*viz.* the inability of the Likert group to obtain university parking stickers because of their nonfaculty status). Even now it is often viewed as a haven of technicians (several steps below the chemistry laboratory staff) currently on many campuses. In contrast there are government and commercial organizations fully devoted to survey design, collection, and analysis. These have decision-making hierarchies constantly monitoring cost and error structures of surveys without the ongoing debate about the relative worth of the enterprise.

The paper ends with a discussion of three developments since 1960 that are important to understanding surveys. At this point, the institutional context is dropped as the organizing principle of the paper and innovations are the focus. Three developments are highlighted: a) the use of the telephone as a data collection medium and later developments in computer assisted telephone interviewing (CATI); b) the use of longitudinal surveys to study micro-level change over time; and c) the application of cognitive psychological concepts to survey methods.

The authors note the movement of mode of data collection from face to face to phone and development of CATI, but they fail to note that this is largely a US phenomenon in the academic and government sectors (the commercial side had done it years ago). Indeed, it is an example of distinctive methodologies pursued by the three sectors. I share their belief that the merits of longitudinal surveys are increasingly being recognized and note that the 1980's is seeing this spread internationally. The Fienberg and Tanur team was instrumental in launching the U.S. effort to apply cognitive psychological concepts to survey measurement, and we are in their debt for this.

The paper does not make it clear whether the authors believe the CATI, longitudinal surveys, and the effort to "cognitize" survey methodological research are the most important three developments in surveys, but they clearly omit several other important ones. We can all choose our three most important developments since 1960; here are some other candidates:

1. Development of Generalized Statistical Software Packages

This development greatly expanded the number of researchers who could directly pose and answer questions using survey data. In the statistical and social sciences at this writing, it is common for undergraduates to perform analyses of survey data whose complexity would have prevented their being done 25 years ago.

2. Existence of Survey Data Archives

The archiving of survey data on computer media was a further democratizing force in survey analysis. With those developments replication and extension of analysis, a key component of the structure of scientific advance, became trivial. Unfortunately, there were also deleterious effects. Analysts of survey data could do their work in complete ignorance of the survey design, of the interviewer training and supervision guidelines, of nonresponse rates, and of a host of other design features known by those conducting the survey.

3. Growth of Commercial and Nonprofit Industry to do Government Surveys

The U.S. is distinctive in its reliance on academic and commercial groups to conduct surveys on behalf of government agencies. Some of this exists in many Western countries, but to a much smaller degree. This suggests that a cross-cultural strain in the paper might be interesting – to identify unique histories of survey research in various societies.

4. 1960 as Beginning of Widespread Acceptance in Academic Circles of the Social Psychological Model of the Interview

This typically describes survey interviews as "conversations with a purpose" and focuses the researcher's attention on the role of the two actors in the errors produced during measurement.

5. Ubiquity of Surveys

Survey measurement is now a way of life for most large corporations (prior to the breakup of ATT in the U.S. the corporation conducted over 7 million customer satisfaction interviews annually). Surveys are viewed as irreplaceable sources of information about customers, suppliers, and the general society.

6. Nonresponse and the Growing Reluctance of the Population to be Measured

This is certainly a phenomenon of great import to survey researchers in most Western countries. With statistical inference to large populations one of the key virtues of surveys versus other data collection schemes, this issue strikes at the heart of the tool. Again, a cross-national theme to the paper would have highlighted these issues.

We can apply the superpopulation metaphor to any historical account – that is, any series of events (which later we call history) is but one realization of an infinite set of possible series which defines the universe of possible realities. This fits the set of questions that remain unanswered.

1. Why after almost a century hasn't survey research fully evolved into a profession (with specified standards and training criteria)?
2. Why is there so little formal educational structure for survey researchers to get their knowledge base? Why are there departments of communications, operations research, naval architecture but none of survey research (teaching sampling, questionnaire design, data analysis)?
3. Would public education about surveys and statistics (like the ASA/NSF program in quantitative literacy) have made an impact on acceptance of surveys?

We are indebted to the Fienberg/Tanur team for reviewing our collective past. They have helped chronicle the birth and first 50 years or so of what is now an important component in most societies of the world. I do hope that the year 2040 will see the need to ask Fienberg and Tanur to update their paper for that occasion. I hope they will be able to report innovation during those 50 years that made a difference in survey methods.

Contributions to Statistical Methodology from the U.S. Federal Government

BARBARA A. BAILAR¹

ABSTRACT

Drawing upon experiences from developments at the U.S. Bureau of the Census, the paper briefly traces some contributions made by practitioners to the theory and application of censuses and surveys. Some guesses about future developments are also given.

KEY WORDS: Sampling; Nonsampling error; Estimation; Confidentiality; Seasonal adjustment.

1. INTRODUCTION

In the United States, the federal government has led the way in the development of statistical methodology in censuses and surveys. I will confine my remarks to examples from the U.S. Bureau of the Census and will discuss four main areas of work – the development of sampling methods, non-sampling error, seasonal adjustment, and the development of methods to protect the confidentiality of respondents, usually called disclosure avoidance techniques. Finally, I will venture to hazard some guesses about future development.

2. SAMPLING

The story of sampling in the U.S. federal government is primarily the story of a remarkable group of people at the Census Bureau, led by Morris Hansen and William Hurwitz. When one considers that the Census Bureau was committed to probability sampling in the early 1940's, one wonders: how could an innovation of this type have occurred so quickly in such a conservative institution? The adoption of innovative methods often takes a very long time and I suspect the Bureau is much slower in adopting and promoting new methodology today. Hansen has given three reasons why he thinks sampling was accepted relatively quickly by the subject-matter divisions of the Bureau. They are: (1) support from the top, (2) conscious development of a team-work approach with the subject-matter divisions, and (3) the development of a corps of sampling experts (later, methods specialists) in the subject-matter divisions who were responsible to the Statistical Research Division (SRD) on technical matters. I think he left out one key ingredient and that is the force and the spirit of the dynamic duo and their cohorts.

In 1936, the Bureau began exploration of sampling and potential applications. Some sampling was already in use, but not probability sampling. There was judgment sampling and sampling of some large establishments. However, there was little or no theory to guide sampling approaches. In 1937, Congress authorized a national voluntary registration of the unemployed and partially employed. A questionnaire was to be delivered by the Post Office to every household. There was some concern that this voluntary registration could have some bias, so an enumerative check census was put in place in a sample of areas. The check census required interviewing all households within a probability sample of postal delivery routes. The mail

¹ Barbara A. Bailar, American Statistical Association, 1429 Duke Street, Alexandria, VA 22314-3402.

carriers did the interviewing and identified and sorted the voluntary mail returns. They then provided separate counts for each postal route, including the sample postal routes. This then gave an independent variable to use in the estimation, one of the earliest demonstrations of ratio estimation. The results of the check census were convincing on the usefulness of sampling. However, the entire effort was remarkable in many ways:

- the effects of nonresponse from a voluntary census were anticipated;
- the use of ratio estimation;
- the speedy results.

Hansen, in an interview in *Statistical Science* (Olkin), reports that the registration took place the week of November 20, 1937; that the household canvas was done during the week of December 4, 1937; and preliminary results became available on New Year's Eve, 1937. I don't think the Census Bureau could beat that record now.

Hansen attributes the success of the 1937 enumerative check census as a demonstration of the use of sampling as key in gaining acceptance within the Bureau. Before then, Bureau staff believed that complete coverage was necessary and that sampling would discredit the Bureau. The success of the study helped gain the acceptance of sampling in the 1940 census, the first census in which some questions were asked of only a sample, not the entire population. Unfortunately, in the last few months, some at Census have dragged out the old chestnut about needing to do the vacant delete check on a 100% basis because a census has less error than a survey. Let's just assume that was a temporary aberration caused by litigation.

A great deal of the theory of sampling was developed in conjunction with the Labor Force Survey. The Works Progress Administration (WPA) sponsored a survey to measure unemployment. In 1942, when the WPA was abolished, the survey was moved to the Census Bureau. The sampling procedures were evaluated and many improvements were made. Several important contributions to sampling theory came from that revision. Some of the sampling principles introduced into the 1942 revision were: enlarged primary sampling units, sampling with probabilities proportionate to a measure of size, and area substratification. These principles were discussed in a 1943 paper by Hansen and Hurwitz in the *Annals of Mathematical Statistics*. Rereading this paper, "On The Theory Of Sampling From Finite Populations," always provides new insights. The article seems to be the first published by federal employees on the topic of sampling of finite populations. Though the concepts had been discussed by others, the extension of theory was new. Also, a hallmark of Hansen and Hurwitz, the results were discussed in a series of practical comparisons highlighting the advantages of the recommended procedures.

Improvements in the Labor Force Survey continued over the years. Composite estimation, using the system of sample rotation to improve the estimates, was introduced. The Current Population Survey, as the Labor Force Survey is now called, has undoubtedly led the way throughout the world in setting the standards for a labor force survey.

Surveys of business establishments presented new sampling problems, also undertaken by the Statistical Research Division. The attitude frequently encountered was that sampling might be all right with relatively homogenous populations such as people but they would not work with highly skewed populations such as businesses. Working with the acknowledged skewness of the population, the sampling group stratified the retail stores by size. The largest stores were necessarily included in the sample, and the smaller businesses were sampled with probability proportionate to a measure of size.

It was also apparent that businesses came into being and died frequently. A static sample would not be able to capture this turnover. Therefore, an area sample to provide estimates

for new stores was incorporated. The Monthly Retail Trade Survey has seen many innovations, but these basic cornerstones remain. The Retail Trade Survey also makes use of composite estimation to provide more precise estimates.

Many other instances of sampling innovations could be mentioned. Many descriptions are given, and the theory and practical applications are described in the book *Sample Survey Methods and Theory* in two volumes, by Hansen, Hurwitz, and Madow (1953). Though the illustrations are seriously outdated, the books still provide more practical sampling applications than any other books I know of. I only regret that they were never updated.

3. NON-SAMPLING ERROR

Another major advance in sample surveys and censuses was to look beyond sampling error to try to control the errors arising from other sources, such as the interviewers, processors, questionnaires, and so forth. Hansen and Hurwitz moved in that direction before the 1950 Census, incorporating many experimental studies in the census designed to estimate the effect of measurement errors in the census. Total survey error became a strong focus at the Census Bureau. The measurement and control of nonsampling errors became a regular feature of Census Bureau work.

An impetus to this nonsampling error work was the recognition that measurement errors could have a much stronger effect on data than sampling errors, especially at larger levels of aggregation. Hansen, Hurwitz and Bershad (1961) developed an integrated model for censuses and surveys that explicitly incorporated sampling error, response error, and bias. The response error component contained what are now known as a simple response variance and a correlated response variance. The simple response variance reflects the basic trial-to-trial variability that arises from differences in respondent reporting, different respondents, different interviewers, and the like. The term has also been generalized to include the variance that arises from trial-to-trial variability in coding. The correlated response variance refers to the variance that arises from a factor that pushes responses into a certain pattern. The most studied factor is that of the interviewer. By having certain expectations or from experience interviewing at a few households, the interviewer can push responses into certain categories. We see wide variability among interviewers working in the same areas on nonresponse rates, on questions about educational attainment, and many other items.

This model was first tested in the 1950 census and was a major factor in the decision to move from an "enumerator census" where an interviewer went to every household, asked the questions, and recorded the answers, to a "mail census", where the questionnaires are sent to every household and householders are asked to fill out the forms and return them by mail. Experiments in the 1960 and 1970 censuses show a large reduction in this variance component when self-enumeration is used (U.S. Bureau of the Census 1968, 1970).

In addition, Hansen and Hurwitz encouraged work on coverage error. The Census Bureau has invested a large amount of time in investigating the effects of coverage error, both in censuses and surveys. After the 1950 census, using a model developed by Ansley Coale at Princeton University, the Census Bureau was able to measure the amount of undercounting in the decennial census at the national level, by age, race, and sex. This method, known as demographic analysis, showed that there was a differential undercount that affected blacks much more severely than whites (Citro and Cohen 1985). In addition, the Census Bureau started development of a post-enumeration survey to learn more about the uncounted population. At first, the Bureau relied on a "do-it-better" approach, but in recent years has turned

to a "do-it-again" approach. This latter emphasis will be used in the 1990 census. Similarly, coverage losses in surveys spurred work on ratio estimation procedures that would dampen the effect. Most Bureau household surveys use those procedures.

The Bureau of the Census now is well known for its work on measurement error. In addition to work on response error and coverage, it has encouraged work on time-in-sample biases that affect the estimates from surveys in which respondents are contacted more than once. The labor force survey, in which respondents are kept in sample four successive months, dropped for eight months, and then contacted for four additional months, has been carefully studied. Bailar (1975) showed the difference between the higher estimates of employment and unemployment for those in sample for the first time and those in sample for later times. These differences affect the levels of employment and unemployment, though probably not the estimates of month-to-month change.

These are only a few examples of the work begun at the Census Bureau on measurement errors. Now work is carried on at all the statistical agencies.

4. SEASONAL ADJUSTMENT

The history of seasonal adjustment in the government began with the efforts of Julius Shiskin when he was at the Census Bureau. He was responsible for introducing computerized seasonal adjustment. Now the X-11 method is used around the world.

According to Julie Shiskin, in the 1950's the Federal agencies were under pressure from the Council of Economic Advisors to produce seasonally adjusted time series. The Census Bureau got the first electronic computer dedicated to data processing, the UNIVAC I, in 1953 and Julie heard a lot about how difficult it was to program from Eli Marks who was in his car pool. It dawned on Julie that the computer could be used for making the seasonal adjustments, so he checked with a computer technician and found that it would take 1 minute to do a 10-year series. Of course, it takes less than that now.

Seasonal adjustment is still somewhat of an art form, since the X-11 program provides so many options, and the analyst can choose among them. However, there was skepticism at the beginning of this computerization about whether a machine could do what a skilled technician could. Julie decided to challenge the Federal Reserve Board. He proposed that they take any series and spend as much time as they wanted adjusting it. Then he would run the same series through the computer. Both series would be plotted and given, without identification of who did the adjustment, to a small, very distinguished group at the Federal Reserve Board who would judge the results. The result was a unanimous decision that the computer method was superior.

The government now seasonally adjusts thousands of time series annually. Model-based methods, because of computer limitations, seemed impractical for many years. Also, new seasonal adjustment factors were developed every year, based on historical experience. For example, a factor to be used in the computation of the seasonally adjusted figures for July would be developed in December of the preceding year. No new data based on more recent events were allowed to influence the adjustment. This made sense when it took several days to prepare punch cards and run the series. But within the last ten years, that method received more criticism and the method of concurrent seasonal adjustment was promoted. The time series staff at the Census Bureau, led by David Findley, did a thorough investigation of the merits of concurrent seasonal adjustment on Census Bureau series, and led the way for the adoption of that method by the Bureau.

The time series staff has also asked some very key questions that are central to seasonal adjustment. First, what kind of standard exists to judge whether or not a series should be seasonally adjusted? Second, given that there are several methods for adjusting time series, how do you evaluate the different methods? In a key paper, Bell and Hillmer (1984) question the need for seasonal adjustment if series can be adequately modeled. They also describe some criteria for evaluating seasonal adjustments. I must be quick to point out that the Census Bureau is not the only government agency that has done ground-breaking work in this area. In fact, one very useful accomplishment of the time series staff at the Census Bureau is to hold regular meetings of interested and involved experts throughout the government. Thus, people at the Federal Reserve Board, Bureau of Labor Statistics, Energy Information Administration, and the Bureau of Economic Analysis, to name only a few, all participate and keep up-to-date on new developments. Estella Dagum at Statistics Canada has led many very successful efforts, including the development of the X-11 ARIMA method.

5. DISCLOSURE AVOIDANCE

Whether or not one agrees with the Census Bureau on its policies about keeping data confidential one must agree that the Bureau has promoted disclosure avoidance techniques to protect data. Disclosure avoidance is an attempt to protect the answers of individual respondents. It has long been a problem in censuses, but is also a problem in surveys, especially surveys that are longitudinal in nature or where records exist that could be linked to the survey results.

Disclosure avoidance problems in the population censuses focus on disclosures that would occur from the publication of very small frequencies. These small numbers lead to the potential identification of single respondents or small groups of respondents. In addition, zeros in cells may also lead to disclosure. Disclosure in frequency tables is usually defined in terms of a threshold rule that states that disclosure occurs if, given any tabulation cell X , one can infer that the number of respondents in X is less than a predetermined threshold value. In 1980 decennial census publications this predetermined threshold value was defined separately for households and persons.

Methods for controlling disclosure in frequency count tables fall into three categories: suppressing all values, perturbing cell values, and replacing numeric cell values by intervals. Cell suppression insures that numeric values are not given and that inferences cannot be derived from manipulation of linear relationships between unpublished and published cell values. Data perturbation means adding or subtracting a small amount from most cell values so that inferences regarding the tabulated values cannot be made with certainty. The third method, replacing point estimates by intervals, is not useful for many data users for cross-classifications.

Cell suppression was the main technique used by the Census Bureau through 1980. Additive restraints along rows and columns of the table generate a series of linear constraints. Once the primary disclosures have been suppressed, mathematical programming is used as a disclosure audit on the table. Though this method was used on an ad hoc basis for years, Cox and his colleagues at the Census Bureau derived the mathematical underpinnings (Causey, Cox and Ernst 1985) and showed how complex cell suppression actually was.

Data perturbation methods, including random rounding, have been developed and used in the United Kingdom, Sweden, and Canada. All of these methods depend on adding or subtracting a small value, sometimes zero, from table cells, with a specified probability.

For data such as sales, value, inventory, and financial information from manufacturing and retail establishments, the Census Bureau is concerned about being able to identify the amount

from respondents. If a competitor reviews a tabulation and subtracts the amount for his firm, the amount for another respondent may be identified. Cell suppression techniques are used. The so-called (n, k) -rule states that X is a disclosure cell if a fixed number of respondents n account for more than a fixed percentage of k of the total cell value. This rule belongs to a class of cell dominance rules, all of which are additive.

Disclosure avoidance work is going on all over the world, primarily in government offices. No doubt this reflects the fact that these offices have serious problems that have been pushed to the fore by the demand for microdata.

6. A LOOK TO THE FUTURE

All four areas presented so far have relied on the development of mathematical models. Sampling, of course, relies on randomization methods, but the control of total survey error led to the formulation of a survey error model, first described by Hansen, Hurwitz, and Bershadt (1961). That model and the experiments used to estimate the parameters were the basis for many policy decisions on the conduct of censuses and surveys.

Time series models are used widely around the world, replacing empirical methods such as the X -11. Researchers are now urging that time series methods become integrated with survey estimation methods to produce more accurate results. It will be interesting to observe how or whether this melding will take place.

Another area of active modeling within government agencies is to produce small-area data. Data are often collected for larger areas of aggregation, such as states, and then data needs are expressed for smaller areas, such as counties. Conferences have been held comparing and evaluating different techniques for producing small-area data. The Census Bureau used empirical methods to develop population estimates during the decade. Several models were explored as part of the undercount research at the Census Bureau, and much was learned about the problem.

Ad hoc methods for editing and imputation are now being carefully scrutinized and mathematical models are being developed. We shall undoubtedly see more modeling of this type in the future.

Thus, the future, as I see it, will be a further expansion of models. This is not to denigrate the empirical methods used now. Statisticians have always recognized that theory and practice go hand in hand. Empirical methods that seem to work lead to modeling and theoretical developments that are tempered by practical experience. The government agencies have many fascinating statistical problems that will lead the way, as they have in the past, in certain areas of statistical methodology.

REFERENCES

- BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BELL, W.R., and HILLMER, S.C. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-317.
- CAUSEY, B.E., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- CITRO, C.F., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census*. Washington, D.C.: National Academy Press.

- DUNCAN, J., and SHELTON, W. (1978). *Revolution in United States Government Statistics*. Washington D.C.: U.S. Government Printing Office.
- HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., and BERSHAD, M.A. (1961). Measurement errors in censuses and surveys. *Proceeding of the International Statistical Institute*, 38, 358-374.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W. (1953). *Sample Survey Methods and Theory*, Vols. 1 and 2. New York: John Wiley and Sons.
- OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 191-210.
- U.S. BUREAU OF THE CENSUS (1968). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders*. Series ER 60 No. 7.
- U.S. BUREAU OF THE CENSUS (1979). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1970: Enumerator Variance in the 1970 Census*. PHC(E) No. 13.

COMMENT

G.J. BRACKSTONE¹

1. Introduction

This paper confirms the significant contributions to statistical methodology made by the Bureau of the Census over the past 50 years. The four examples chosen by Bailar to illustrate these contributions are striking, not only in their intrinsic importance, but also in their variety. These are not variations of a single methodological breakthrough; they are fundamental contributions in four distinct areas. They are perhaps themselves illustrative of the wide variety and challenging nature of methodological problems faced by government statistical agencies – a variety and level of challenge that belie any suggestion that government statistics involves only the routine and the mundane.

Of particular interest in the description of these examples are the insights into the environments in which these developments came about. While the methodological contributions have themselves yielded benefits far beyond the original problems they were designed to address, the processes that led to these original contributions are themselves worthy of attention to identify the circumstances that need to exist to make such breakthroughs possible. I will return to this theme below.

During this same period, the Bureau of the Census was also making significant contributions to the automation of statistical processes. Having pioneered the development of punched card sorting and tabulating equipment in the earlier part of the century, the Bureau of the Census was responsible for the introduction of the first computer into a statistical agency in the 1950s. Subsequently in the 1960s, the Bureau also led the way in the automation of data entry by developing FOSDIC, a device for reading a microfilm copy of a marked questionnaire. Clearly the innovative contributions of the Bureau of the Census permeate many aspects of the work of government statistical agencies.

2. The Diffusion of New Methodology

Each of the contributions to statistical methodology described by Bailar originated with a real practical problem faced by a statistical agency. The need to collect additional data at reasonable cost and with acceptable timeliness motivated the development of sampling methods; the need to improve data quality by understanding, measuring and reducing non-sampling errors led to work in this area; seasonal adjustment developments seem to have been prompted by a need to speed up and standardize a skilled manual procedure; the problem of defining a rational and efficient process for ensuring the confidentiality of individual information in statistical outputs inspired the research on disclosure avoidance. Each of the many other examples that could have been cited share this characteristic of having had a real practical problem as catalyst.

The successful development of statistical methodology to address problems such as these is clearly of direct benefit to the statistical agency involved. But have these contributions had benefits more broadly? Have they added to the body of knowledge and methodology known as Statistics? It will be argued that these developments have had significant and broad benefits to statistical agencies engaged in the production of social and economic data, but that their impact on the subject of Statistics as treated in universities, while growing, has not been as influential as it might have been.

¹ G.J. Brackstone, Assistant Chief Statistician, Statistics Canada, Ottawa, Ontario.

Firstly, consider other government statistical agencies. In most countries the government statistical agency is a unique organization dealing with the problems of running regular large household and business surveys, integrating data from various sources, maintaining and analyzing time series, and making large volumes of data available to the public. (In this respect the United States is an exception in having several major organizations involved in this type of activity in different subject areas.) In most countries, therefore, statistical agencies have to look abroad for experiences similar to their own and for peer discussion and review. The network of interaction between government statistical agencies is extensive among developed countries. Contacts may be bilateral or multilateral. The long-standing and continuing exchange of information and experience between Statistics Canada and the U.S. Bureau of the Census is an example of the former. Statistics Canada has benefitted greatly from being able to adopt, and in some cases extend, statistical methodologies developed at the Bureau of the Census, including all of those described by Bailer; equally, I believe, the Bureau of the Census has benefitted from methodological developments at Statistics Canada.

On the multilateral level, several organizations provide regular fora for the exchange of information between statisticians in government agencies. These include the United Nations and its regional and specialized bodies, the International Statistical Institute, particularly its sections for Survey Statisticians and Official Statistics, and the professional statistical societies of several countries. In addition, both U.S.B.C. and Statistics Canada have instituted annual symposia or research conferences at which new developments and experiences are exchanged. All in all, this mixture of bilateral and multilateral contacts serves well to ensure that contributions to statistical methodology emanating from any agency – and many agencies are making significant contributions – are freely shared and utilized in other agencies.

But what has been the impact of such developments on the statistical profession outside government statistical agencies? Here we will use the specific examples cited by Bailer for illustration, though there are many other areas (some of them listed in Section 4) for which similar arguments would apply. In the case of sampling, the influence on the profession has been far-reaching. The topic of sampling from finite populations is now an established part of many university statistics curricula and is the subject of numerous textbooks. The developments initiated in a government statistical agency have been absorbed and extended by the profession. Indeed, some might argue that they have in some respects been taken far beyond the practical needs of survey-takers. In the case of non-sampling errors, the story is different. These developments have not yet led to a well-established body of theory and methods. That is not to say there have been no developments. On the contrary, there has been a wealth of work. However, much of it has been survey specific. It has improved, one hopes, many individual surveys, documented a great deal of experience, and generated a certain amount of applicable wisdom. But the topic has not yet found a secure niche in statistics curricula. Indeed, the accrued wisdom is often associated with particular areas of application (sociology, demography, *etc.*) rather than with Statistics as a subject.

Seasonal adjustment provides yet another story. With its origins as a rather empirical process used in statistical agencies, it has attracted increasing attention in recent years with attempts to provide it with a sound statistical basis. Bailer refers to some fundamental questions about objectives and yardsticks for seasonal adjustment that are now being addressed. Model based alternatives to the traditional X11 approaches are also being investigated. This is an area of statistical research that has attracted attention among time series experts in universities. Seasonal adjustment techniques clearly have applications well beyond government statistical agencies.

Finally, the most recent example that Bailer describes is disclosure avoidance. This is a problem largely confined to agencies operating under a confidentiality code that prohibits

divulgence of any identifiable individual information. Most of the research in this area is taking place in statistical agencies. The tools being used, however, tend to be from the fields of computer science, numerical analysis and mathematics. This is a relatively new field that has not yet attracted much attention outside government statistical agencies.

These examples show that methodological contributions from government statistical agencies not only solve problems for these agencies but can also lead to significant advances in the field of statistics more generally. Of course, not all such contributions have wide applicability and some may remain confined essentially to statistical agencies. A continuing challenge for government statisticians is to generate interest among other statisticians, particularly those in universities, in research problems arising in government work.

3. An Environment for Innovation

Innovative contributions rarely arise by chance. A suitable environment that allows ideas to develop and research to flourish is required. This is not always easy within an organization whose primary mission is the regular dissemination of data according to pre-determined schedules. Bailar refers to three reasons given by Hansen why sampling was accepted relatively quickly in the Census Bureau. In essence, these same three reasons define prerequisites for an innovative research environment in a statistical agency:

- (a) management support in the sense of a willingness to invest in research activity;
- (b) co-operative clients in the sense that successful research needs a particular application that represents the initial problem and sets the research schedule – the manager of this program has to be an enthusiastic guinea pig;
- (c) competent research staff, not just in terms of expertise in particular areas, but also in terms of the ability to recognize problems susceptible to generalization and solution through statistical methodology.

While these three conditions will help to provide an environment conducive to research, further effort may be required to ensure that research results are in fact used, and used appropriately. This requires persuasiveness and good communication skills on the part of the statistician, as well as adequate institutional support for the new methodology.

4. Other Contributions

Bailar was not trying to be exhaustive in her examples of contributions to statistical methodology. It is worth noting some other areas of statistical methodology in which statistical agencies have made significant contributions. Some of these are mentioned as future topics by Bailar, but pivotal contributions have already been made. The following areas would find a place on a Statistics Canada list.

- (a) **Methods for analyzing data from complex surveys** Of great relevance to users of most government statistics, these methods aim to adapt or replace traditional methods of statistical analysis that assume simple random sampling. This is an area of work that has attracted the interest of university researchers who have also made many contributions to the topic.
- (b) **Record linkage** This technique is used in deriving statistics from administrative records, in micro-matches to assess quality, and in list frame maintenance. The development of a general theory for record linkage has provided a basis for software to support this activity. Most of the work on this topic has emanated from statistical agencies.

- (c) **Editing and imputation** Widely used in many surveys, this technique lacked a sound statistical basis until theory was developed in the 1970s. Since then methodologies and systems have been developed to provide general facilities for performing these functions in a variety of surveys. This topic has generated substantial interest and further work outside statistical offices.
- (d) **Small area estimation** In recent years the production of estimates for areas smaller than could be supported by direct estimation from sample surveys has received increasing attention. Statistical offices have developed a variety of methods to address this problem and university researchers have participated actively in this work. To date the utilization of such methods for production purposes has been limited, partly due to lingering concern about the probity of government agencies producing model-based estimates.
- (e) **Statistical use of administrative data** As another means of reducing data collection costs, the statistical potential of existing administrative records has been exploited. Such sources present a different array of coverage and data quality problems, from those experienced in surveys. While administrative data may be used alone to produce statistical data, they may more effectively be used in combination with survey or census data in estimation systems that take advantage of the relative strengths of each. Most of this work has taken place in government statistical agencies.

5. Future Areas

In looking to the future, Bailar foresees increased use of models. This is almost certainly correct as statistical agencies strive to extract the maximum information out of existing data and minimize the increasing costs of data collection. In particular, she refers to the melding of time series methods with survey estimation methods, an area now being explored in several statistical agencies. I would add three other domains of activity in which we might look forward to significant developments in the long run, each of them requiring an interaction of statistics with other disciplines.

The first is the application of expert systems to certain activities in government statistical agencies. To use an example already discussed, the choice of the appropriate options or models to use in seasonally adjusting a time series could well lend itself to such an approach. The second area is the use of cognitive methods for understanding and improving the response process. Work in this area is underway at a number of statistical agencies. Drawing on the expertise of psychology, it may provide a basis for enabling statisticians to develop better models of the response process – probably the least well understood component of the survey process. The third area is the development of integrated statistical information systems that combine models of social or economic systems with databases on which the impact of different policy assumptions can be simulated. Such systems serve to facilitate the use of an agency's data for policy analysis, and also help it to recognize data gaps in current programs.

To echo Bailar's conclusion, the problems are fascinating and there are more than enough to go around.

Rolling Samples and Censuses

LESLIE KISH¹

ABSTRACT

Rolling censuses combine F nonoverlapping periodic samples of $1/F$ each, so designed that cumulating the F periods yields a complete census of the whole population area with $F/F = 1$. Intermediate cumulations of k samples would yield samples of k/F for more timely uses (annual or quinquennial censuses). Area sampling frames would cover the national territory for naturally mobile populations. These methods may often be preferable to other alternative methods for censuses, also discussed. *Asymmetrical cumulations* are also recommended to counter the problems of small sample cells for area domains (provinces, regions, states) common to most countries and to other population units. *Split-panel-designs* offer another use for cumulating periodic surveys by combining nonoverlapping portions $a - b - c - d -$ with panels p for partial overlaps, $pa - pb - pc - pd -$, for multipurpose designs.

KEY WORDS: Periodic samples; Time sampling; Cumulations; Split-panel designs; Asymmetrical cumulations; Multipurpose designs.

1. INTRODUCTION AND DESCRIPTIONS

Several uses and methods for cumulating data from periodic samples are discussed below. This has been a rather neglected subject, as the literature on periodic and rotating samples has concentrated on the statistics for net changes and for current ("cross section") estimates; not on cumulations. The first concern here is on rolling censuses and samples, and let me attempt a definition of *rolling censuses*: a combined (joint) design of F separate (nonoverlapping) periodic samples, each a probability sample with fraction $f = 1/F$ of the entire population, so designed that the cumulation of the F periods yields a detailed census of the whole population with $f' = F/F = 1$. Intermediate cumulations of $k < F$ periods should yield rolling samples with $f' = k/F$ and with details intermediate between 1 and F periods. We may appreciate that definition by looking at examples and counterexamples. We shall also examine possible variations that would satisfy the definition and conflicting needs that rolling samples can be aimed to meet.

Imagine a weekly national sample, each with *epsem* selection rates of $1/520$, and so designed that in 520 weeks they are "rolled over" the entire population and the cumulation yields a complete census of the population averaged over ten years. Each year would yield national and local samples with selection rates of $52/520 = 1/10$. The design would combine weekly national samples into an averaged decennial complete census, and into sample censuses of ten percent each year.

The Health Interview Surveys of the National Center for Health Statistics (1958) cumulate 52 weekly samples of about 1,000 households each. These samples select about $f = 1/80,000$ weekly; thus $520/80,000$ represents cumulations of nonoverlapping periodic samples over ten years. But they are confined to a set of PSU's for reasons of cost chiefly, but also for better estimates of net change and for current estimates. However, *rolling samples* may better be reserved for samples designed for maximizing (increasing) the spread (representation) of the samples cumulated over national (or broad) populations. The words in the parentheses indicate that rolling samples constitute a special case of the more general *cumulated periodic samples* and that the boundary of the subset need not be precisely clear.

¹ Institute for Social Research, The University of Michigan, Ann Arbor, U.S.A. 48106.

For overlapping between periodic surveys, the requirements for the selection of units of cumulated designs are diametrically opposed to the requirements for the objectives and substantive content of the interviews (the observations, variables). The content of the surveys must be as similar, standardized, identical as possible for the cumulations to be meaningful. Using periodic panels of the same elements for different contents could broaden the scope of surveys, but would not contribute to increasing the sample size for survey statistics. Most periodic surveys collect similar variables, though some may also have other contents attached at times. However, changes of methods, questions, and variables would cause conflicts and problems. Perhaps such changes should be introduced only with extended intervals of "splicing", using both the new and the old methods to study the differences. These problems are fundamentally similar to those faced when measuring differences from periodic surveys, but they seem more novel. I insist (Section 6) that solutions to such problems must be tailored to specific situations.

On the other hand, the cumulation of the same elements (persons, households) does not increase proportionately the sample size (base), and panels of the same elements would not help rolling samples. Many periodic surveys (*e.g.*, labor force surveys of Canada, the USA *etc.*) have partly or largely overlapping fractions of segments (ultimate clusters), and those tend to contribute little toward increasing the sample size. Even in surveys with nonoverlapping segments (like the HIS of the NCHS (1985)), the segments are confined to the same first stage (and second-stage?) units; in these the positive correlations (clustering effects) tend to reduce the "effective" sample sizes for overall statistics. Furthermore, those periodic samples, confined to samples of primary units fail to meet the needs of rolling samples for spreading over the entire (national?) population.

A few more remarks may help to broaden our frame of reference. (1) The discussion often assumes area sampling, but the concept can be generalized to other frames. (2) Equal selection rates for elements are often used, but cumulations may be modified to unequal selection probabilities. (3) The concept may be generalized from regular periodic samples to cumulations over less regular periods. (4) Cumulations over the entire time span (year or ten years) come most readily to mind, but we may envisage systematic sampling of the span; *e.g.*, labor force surveys cover only single weeks of the months over the year.

2. ALTERNATIVE METHODS FOR CENSUSES

Rolling censuses would be expensive, and the reason for such an innovation should include the acknowledged relative weaknesses of the decennial censuses now widely used, and of sample surveys and administrative registers, which are proposed at times as possible alternatives. The chief reason for censuses is the need for detailed information, especially for small areas; and the chief weakness of decennial censuses is their obsolescence between censuses and their great total cost that prevents more frequent censuses. Sample surveys have many advantages for national statistics and for large regions, but they lack geographical and other details. Good registers are rare and they provide few variables beyond a few, bare demographic data.

Decennial censuses of population, housing, agriculture, industry and others, first and foremost, have spread into most countries in the last two centuries, and especially in the last two generations with the help of the United Nations State Statistical Office. In addition to detailed data for small domains, censuses often may obtain better coverage than samples, due to the concentrated publicity and the national "ceremony" connected with censuses; the Chinese census of 1982 is a good example (Kish 1979, 1989). The efforts of the census also yield *lower unit costs* (for short forms) than surveys, but much *higher total costs* than sample surveys, because of much greater size. At 2.6 billions, the 1990 censuses of the USA will cost \$10 per

capita or \$30 per household. That cost of about half to one hour of the median hourly wage per capita (once in ten years) seems to hold in international comparisons, though the number and complexity of census variables is one of the cost factors. Rolling censuses would probably be proposed and designed for surveys fairly rich in the numbers and complexity of variables. In Canada 260 weekly samples of 32,000 households would cumulate to the national population. In the USA 520 weekly samples of 160,000 would be needed by decennial cumulations to 80,000,000 households; the CPS surveys have 100,000 with state supplements.

No detailed comparison of decennial censuses with rolling censuses is possible here, but the issue of *timeliness* must be mentioned, because that is the chief issue in the comparison. Up to now the periods for using data from decennial censuses have varied from a start of 1-4 years to 14 year or more. Even with faster computers the start is slower for complex social statistics than for mere head counts; and the obsolescence over the ten intercensal years becomes worse with higher population mobility in our modern civilization. The biases due to obsolescence will be monotonic, if not linear, functions of elapsed time. The sizes of the biases will differ with variables, populations, etc.; but they will be present and considerable, I believe; often perhaps greater even than the famous biases due to under coverage (Kish 1981, 1979).

Increasing and rapid obsolescence of decennial census data should chiefly motivate the searches for alternatives, such as in *A Study on the Future of the Census of Population: Alternative Approaches* (Redfern 1987). "A serious weakness of the census is that it occurs relatively infrequently". About a "rolling census" it states: "The merit of this proposal is that . . . a much smaller, better trained organization and more experienced staff could be deployed both for the fieldwork and for processing . . . the public awareness of the rolling census would not be highly peaked. Whilst that might well lessen the risk of public protest, the reduced publicity would adversely affect the level of coverage achieved . . . (The method) would complicate the interpretation of the census results, especially comparisons between areas. Simultaneous national coverage, one of the virtues of the census, would be lost. The idea of a rolling census has not yet been developed and applied".

Most countries will probably still need censuses in 2000 AD. They are being replaced by population registers in the Nordic countries and still need to be introduced in some Third World countries in 1990. They have been stopped by opposition and by obstacles in a few. But most countries need and will have them in 1990. They have been a great and useful invention – like the steam locomotive, and at about the same time. However it is possible that the censuses also may be phased out gradually by some of the alternatives here considered.

Quinquennial annual censuses have been proposed, and quinquennial censuses have been initiated or carried out in a few countries, including Canada and Turkey. But these are not destined for quick acceptance, I suspect. They seem too costly: ten percent samples in two countries had half of the costs of complete censuses. Also they still leave a great deal of obsolescence. On the other hand, much smaller (e.g., 5 or 1 percent) yearly sample censuses would fail to offer enough geographic detail. The one percent "microcensus" of West Germany provides yearly sample data. China had a one percent census in 1987; their yearly samples of 1/2,000 (also about 500,000 people) collect chiefly fertility data only (State Statistical Bureau 1987; Kish 1989). Quinquennial censuses are not frequent enough and yearly censuses would be too costly.

Administrative registers provide a great deal of diverse data in many countries, and they are likely to spread in the future. Excellent *population registers* exist in the Nordic countries of Sweden, Norway, Denmark, and Finland, and perhaps in some other countries of Northern Europe. Their completeness is based on cooperation, motivation (with social incentives), and literacy; in a few cases they are replacing censuses with data from the population registers. In other situations their coverage, quality, and updating are far from adequate. We can expect

future improvements in the quality, spread, and use of population registers but not quickly and not widely. We should not expect them to replace censuses even in developed countries like the USA and Canada, and their use in less developed countries soon is even less likely (Redfern 1989).

Furthermore, even after population registers become adequate in quality and coverage, they will contain and supply only a few, bare demographic variables: head counts, age, sex and little more. Thus, they will fail to meet the demands of modern society for richer sources of statistics. For these the registers will serve only as auxiliary variables.

Synthetic, ratio regression, and raking estimators are being used increasingly for *small area statistics* (Platek *et al.* 1987; Purcell and Kish 1980). Census data are usually obsolete, data from registers inadequate, and sample data lack details for small areas. The weaknesses and strengths of the three methods are complementary, hence combining the advantages of the three methods seems like good strategy. This is the common purpose of the several methods of *small area estimation*: to provide estimates for small areas and for other small domains that are current, accurate, and relevant.

These methods are now being used for local area estimates of population counts for the intercensal years, in order to compensate for the obsolescence of the decennial censuses, thus sometimes called *postcensal estimates*. They also have other uses in increasing numbers, *e.g.*, they have been proposed to compensate for undercount biases. However, those methods have all combined censuses with sample surveys and registers. Therefore, they should not yet be considered as alternatives to censuses. Nevertheless, we may raise the question whether rolling censuses would perform better or worse overall than decennial censuses in those combinations. The answer is uncertain, but I believe that the balance of variance components would favor rolling censuses in most cases. However, theoretical as well as empirical investigations will be needed to decide this question as well as several others here.

Partially overlapping samples from multipurpose designs must be considered because they exist in many countries for several purposes and they absorb some of the funds available for national statistics. These multipurpose surveys often provide labor force statistics and other valuable data. They vary in parameters between countries but they also have several basic features in common with those of the USA and Canada. They are periodic samples with overlaps that are constant and for fixed periods (but all three parameters differ between countries). They use area segments for bases, but not panels of households (movers are not followed). The overlaps are usually large and these are generally justified with references to reductions of variances from positive correlations in the overlaps. But an even greater advantage of overlaps may be the lower costs of interviewing in later calls, especially where telephone calls follow first calls on foot. These "rotation designs" have dominated practice and literature and they represent an important innovation (by H.D. Patterson 1950 and R.J. Jessen 1942). They are designed for measuring net changes and current (level) statistics, but not for cumulations. However, the variances (per household) would not be greatly increased for overlaps of even a small fraction (< 0.3), when compared to the large overlap (> 0.7) commonly used. This is particularly true for many variables like being unemployed, which have low correlations between periods. Furthermore the overlaps could be changed in other ways (Section 5). Therefore it is possible that these surveys could be combined with the cumulations needed for rolling samples and censuses.

3. CUMULATIONS OVER TIME AND SPACE

Changes in populations and in their variables are often recognized as of three kinds: "secular" trends, which are more or less smooth and monotonic, like "growth"; periodic and "cyclical", such as seasonal fluctuations; and irregular variations which are difficult to describe

and often treated as "random". Designs for cumulating, averaging, and sampling over temporal variations face psychological obstacles that differ from our acceptance of designs for variations over spatial variations. Spatial variations can be large and sometimes accountable, but more often irregular. However, we have learned to accept samples, averages, and cumulations over them in population (national) aggregates and averages.

The psychological blocks still facing rolling samples and censuses may be countered with both theoretical and pragmatic arguments. The theoretical and philosophical arguments are hinted at above and in later discussions of alternatives (Kish 1987, 6.1B). The pragmatic and empirical arguments may be buttressed with several types of uses we recognize as common and successful. The same periodic samples for obtaining current data and for measuring changes can also be used for aggregates needed for spatial and domain details. Furthermore, by averaging (over a year or longer) the temporal variations (seasonal or cyclical or erratic) are smoothed over in the moving averages.

Retrospective data. "Children ever born" to women who completed fertility over the entire fertile span of 30 years may represent an extreme for retrospective spans; but other individual interview data aggregated over life spans include serious diseases, education, *etc.* Interviews aggregated over yearly spans include farm production, work history, income, home and auto purchases. Of course, all these data have imperfections, which differ across variables, respondents, methods, *etc.* But even cumulations over a week or over a day (such as purchases of bread or cigarettes) have errors. *Multiround surveys* are used for cumulating short term data; for example, births during the past month have been cumulated from 12 monthly samples over the year.

Cumulating rare elements from periodic surveys has often been used to deal with these difficult and expensive problems. The topic has been dealt with and illustrated in publications on rare items (Kish 1965 11.4; Kalton and Anderson 1986). *Statistics for small domains* may also benefit from cumulations, and single years of birth may exemplify such small domains, which consist of "crossclasses". But geographical and administrative units are "proper domains"; for these the periodic samples are not adequate, because those domains need the designs of rolling samples or censuses.

Cumulations from periodic samples. The Health Interview Survey (NCHS 1958), described above, may be the best known example with yearly cumulations of weekly samples of about 1,000 households from nonoverlapping area segments. It is designed for *multipurpose* objectives (like most periodic surveys) including cumulations for some rare diseases, but also estimates of current levels and net changes. It provides some estimates for larger domains, as well as national estimates for the common diseases. To convert it into a rolling sample, by increasing the spread of the yearly samples, would increase field costs, especially in that portion (about 30 percent only) where the PSU's are counties (not self-representing).

A traffic survey provides an interesting example of cumulations, because the population is very mobile within the sampling frame of sampling units of locations x hours (Kish, Lovejoy and Rackow 1961). The general concept is applicable to nomads and other mobile populations. It may also serve less mobile general populations over a longer period, such as the decennial spread.

The earliest cumulation I found is for a sample of California in 1952 (Mooney 1956). "The samples were selected in such a manner that they resulted in a uniform overall sampling rate of 1 in 385. For purposes of enumeration, the sample was divided into 52 equal subsamples, and a different subsample was enumerated during each week of the survey year. Consequently, each week's enumeration was based on a sample of 1 in 20,020". For smaller states (populations) and/or larger samples one may imagine weekly samples of $1/520$, and complete rolling samples in the 520 weeks of the decennial census period. It is likely that such rolling samples have been designed for smaller populations.

The above examples refer to nonoverlapping periodic samples. Cumulations from partially overlapping samples have been used, but with the "effective sample sizes" reduced by the amount of the overlap (Ericksen 1974). Furthermore, this paper concerns cumulations of individual cases, but periodic or repeated surveys may also be used for *combining statistics* from them (Kish 1987, 6.6) as in "meta-analysis".

4. ASYMMETRICAL CUMULATIONS

This term denotes a proposed method of cumulation for problems that arise because "natural" subpopulations generally vary greatly in size. For example, I have been faced within the past few years with ranges of 50 or even 100 to 1 among the provinces (or states) of Canada, USA, Australia and China; and those ranges of relative sizes are similar for the provinces of most countries. Those inequalities arise because administrative units tend to be created roughly equal in areas, but spread over lands with highly unequal population densities. They also exist for districts, counties, *etc.* within most provinces. They also arise for other social units and social organizations, like firms, hospitals, universities. But not for all: military units, census enumeration districts and elementary schools are created roughly equal.

For many other frequency distributions rough equalities of classes are created with traditionally accepted cumulations over roughly logarithmic scales; *e.g.*, income, city size, *etc.* are often tabulated in classes like 10-25, 25-50, 50-100, 100-250, 250-500, 500-1,000, 1,000-2,500, *etc.* This shows a sensible method of cumulation that creates roughly equal cells on a roughly logarithmic scale, and they are traditionally accepted and understood, although highly asymmetrical.

Note also that cells in tables for sample data are *generally cumulated over both space and time*. For example, monthly surveys of labor force often show labor force statistics cumulated over the month (or over a week as a "sample" of the month), and also over the provinces (from a sample of sampling units). Quarterly and yearly statistics show further cumulations, as do the national statistics. The spans of cumulations must balance three parameters of restraints: the span of the reference period that may be relatively flexible; the domains of subpopulations, which may be more rigid, like provinces; and the sample size expressed in sampling units and variance components. Other variables, such as cost factors and "required precisions", tend to be expressed through the basic three parameters of cell size.

Decennial censuses of the population counts represent extremes by emphasizing locational detail: persons are placed in homes as of the reference date (April 1 in the USA). But yearly and longer cumulations are possible for income, *etc.* Time gets sacrificed in obsolescence, and sample sizes and costs in complete coverage. At the other extreme are monthly sample surveys for labor force and health variables, and myriad other variables, where the emphasis is placed on timelines and reduced costs, but at great sacrifice of spatial detail.

Population inequalities between provinces impose severe restraints on timeliness and sample sizes. Often higher sampling rate are introduced for the smaller provinces, but such "optimal" selection rates bring disadvantages in increased variances both overall and for cross-provincial "crossclasses" (age, sex, *etc.*) (Kish 1988, Section 5; Trewin 1987). Thus those mildly unequal rates fail to solve conflicts in provincial sizes of 50:1 or 100:1.

Because of those conflicts the tables for monthly surveys commonly present cells for small provinces with inadequately small sample sizes. Two alternative procedures have been advanced and practiced for such small cells. A. Release the same data for small cells as for large cells, and let the reader (user, consumer) beware, *caveat emptor*, with perhaps warnings posted

to appendixes to sampling errors. B. Don't release, but suppress small cells, leaving them blank, after applying some declared curtailing limits. Readers may be directed to other released publications, based on cumulated data (quarterly, annual).

Asymmetrical cumulation proposes a compromise between symmetrical releases (A) and asymmetrical suppression (B).

C. Asymmetrical cumulation proposes to release for small cells the *specified* cumulations of periodic data. These cumulations may be flexible: for example, quarterly for small cells and yearly for very small cells, instead of the monthly data for large cells. The readers may be notified (with * or italics or other signs); thus they may choose either C (cumulation) or B (disregard).

AC. This procedure would allow readers to choose either A or B or C by publishing *both* the current monthly data A and the cumulated C data.

Procedures B and C have the disadvantage that the cells do not sum to the marginals. But AC like A do sum to the marginals. Some iterative method could overcome these disadvantages of B and C.

5. MULTIPURPOSE SPLIT PANEL DESIGNS (SPD)

In order to find adequate funds for rolling samples and censuses it is desirable to consider how they could be combined with the periodic surveys now being funded and conducted in many countries. These are either monthly or quarterly surveys (sometimes yearly or weekly). They are typically partially overlapping samples designed for improved estimates for current level and net changes. However they are not designed either for cumulated rolling samples, or for panel studies based in the overlaps. I proposed SPD as the design for providing data for all those four purposes; and also for some fringe benefits (Kish 1987, 6.5).

a. Combining two separate periodic samples forms the basis of SPD: to add a panel p to a parallel series of nonoverlapping samples $a - b - c - d$ etc., with the combination then denoted as $pa - pb - pc - pd$ etc. The panel p provides individual (micro) changes and the nonoverlaps can be cumulated into larger samples and rolling samples. The combined samples provide the partial overlaps best for current estimates and for net changes; thus they can replace the usual rotating samples. This combined use is a main feature of SPD, together with the provision of a flexible and potentially large sample of nonoverlapping portion for use in cumulating samples.

b. The designs for p and for $a - b - c$ can be separate and distinct, each "optimized" for its own objective. But they must also be combined for joint estimates of net changes and current levels; and for that purpose the populations covered and the measurements used must be similar enough for the combination.

c. SPD has considerable advantages because its overlaps exist for *all* periods, whereas they are rigidly fixed in classical rotation designs. This advantage is clear and important for net changes because it exists for all desired comparisons. But it also exists for current levels, because the correlations may differ among variables.

d. Including proper panels p of elements necessary for measuring individual (micro or gross) changes would be a great advantage for SPD over partial overlaps now used. However, the other features can be satisfied with overlaps p' of area segments as at present. Furthermore a modest and slow rotation can be built into the design of either the panel p or the overlap p' , so as to retain most of the gains from covariances and from panel information. Perhaps some alternation may be introduced to reduce panel fatigue or deterioration. Several surveys have used *both* the overlap p' and panel p by following as many movers as possible. Most

households belong to both samples. The extra cost for the panel depends on the proportion of movers and their cost (Kish 1987, 6.2, 6.4).

e. The advantages and problems of panel interviewing pose difficult problems, with a large and varied literature and conflicting results (Kish 1987, Sections 6.4, 6.5). The number and spacing of reinterviews that are possible, desirable, and reliable need to be established.

SPD has an advantage in separating the panel p whose cumulated data may be checked against the nonoverlaps for “panel biases”, and perhaps even for adjustments of biases when those are measured adequately.

Another useful modification may be to recruit sampling units into the panel by different (“optimal”) selection rates on the basis of their being “screened” in the nonoverlaps.

f. The size of $a - b - c - d$ need not always be the same; this flexibility of SPD, which differs from the rigidity of rotating designs, may be used for needed sample enlargements or for cost retrenchments. Such changes would raise weighting problems (solvable) for cumulations.

g. The relative size of the panel p against the nonoverlap $a - b - c - d$ portions depends on feasibilities and costs and needs study (Section 6). For individual changes we need larger p , but for cumulations larger $a - b - c - d$. The larger p portions now common may be favored by lower field costs for telephone reinterviews.

Lower values of p than are now common are good enough for current levels and for net changes with weighted estimates; the optima are insensitive and p between 1/4 and 1/2 are all nearly best; lower p may also be used where the emphasis lies in nonoverlaps $a - b - c - d$ for cumulations.

6. CONCLUSIONS AND QUESTIONS

Cumulated samples provide the bases for four new methods proposed here: rolling samples, rolling censuses, asymmetrical cumulations, and split panel designs. Rolling samples have been designed, but the other three still await practical applications. Meanwhile we should welcome methodological developments that would outline the parameters of feasibility.

However, the chief tasks for these methods must be found in the details of specific situations rather than in theoretical generalities. The factors of costs, variances, biases, feasibilities, and public acceptance for novel procedures must be worked out specifically for each situation. We can do no more than raise a few questions as examples, in addition to those raised implicitly or explicitly in the preceding sections.

1. For rolling samples and censuses what kinds of moving averages may prove most useful? For national aggregates the latest month (or quarter or year) may receive the full weight. But for small local areas the data may be cumulated over ten years; with equal or with increasing weights? Are “shrinking” (Stein-James) estimators useful?

2. How to deal in the aggregates with changes in the population, in methods, in variables?

3. For asymmetrical cumulation similar questions arise. Should the latest monthly estimates (A) be printed together with the cumulated (C)? Methods are needed to make the cells and the marginals consistent.

4. For the split panel design, how large should the overlap (p) be? Can it be a panel or merely overlapping segments? Or must we, can we, have both? How does it depend on the correlations for diverse variables? How do we balance the four chief purposes of periodic surveys?

There will be other interesting questions but this essay must come to an end before they do.

REFERENCES

- ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-75.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304.
- KALTON, G., and ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society (A)*, 149, 149-52.
- KISH L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH L. (1979). Samples and censuses. *International Statistical Review*, 47, 99-109.
- KISH L. (1981). *Using Cumulated Rolling Samples*. Washington: Congressional Research Office, 80-528-0.
- KISH L. (1987). *Statistical Designs for Research*. New York: John Wiley and Sons.
- KISH L. (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.
- KISH L. (1989). Developing statistics in China. *Journal of Official Statistics*, 5, 157-69.
- KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A multi-state probability sample for traffic surveys. *Proceedings of the Section on Social Statistics, American Statistical Association*, 227-230.
- MOONEY, H.W. (1956). *Methodology in Two California Health Surveys, San Jose (1952) and Statewide (1954-55)*. U.S. Public Health Monograph No. 70.
- NATIONAL CENTER FOR HEALTH STATISTICS (1958). *Statistical Designs of the Health Household Interview Survey*. Washington: Public Health Series, 584-A2, 15-18.
- PATTERSON, H.O. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 16, 140-149.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: Wiley-Interscience.
- PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas. *International Statistical Review*, 48, 3-18.
- REDFERN, P. (1987). *A Study of the Future of the Census of Population: Alternative Approaches*. Luxembourg: Statistical Office of European Commission No. ISBN 92-825-7429-6.
- REDFERN, P. (1989). Population registers: some administrative and statistical pros and cons. *Journal of the Royal Statistical Society, A*, 1-41.
- STATE STATISTICAL BUREAU (1987). *The 1987 Nationwide One-percent Population Sample Survey*. Beijing: State Statistical Bureau.
- TREWIN, D. (1987). Estimation of trends and time series models from continuing surveys. *Bulletin of the International Statistical Institute*, 46.
- UNITED NATIONS (1981). *Principles and Recommendations for Population and Housing Censuses*. Series M No. 67 (E.80.XVII.8).
- REDFERN, P. (1989). Population registers: Some administrative and statistical pros and cons. *Journal of the Royal Statistical Society, A*, 152, 1-41.

COMMENT

FRITZ SCHEUREN¹

The statistical literature has neglected the idea of cumulative samples. Leslie Kish, in several previous papers and in the present one, has tried to rectify matters. Ever forward-looking and practical, he makes a persuasive and compelling case for more work on the design and analysis issues raised by cumulation.

His writing is so down-to-earth that readers may miss the fact that Kish is not just advocating a few minor additions to the already large supply of survey designs and estimation methods. He asks us to look very hard at the topology of the space/time/content trade-offs in surveys – especially in censuses. In fact, Kish seems to be advocating what might be called a “paradigm shift” in census-taking, at least in developed countries like Canada and the U.S.

The word “paradigm” deserves some elaboration (Barker 1988). A paradigm is a way of thinking and then doing, a pattern of belief and behavior, a way of seeing reality and using that sense to accomplish something. Paradigms are common – the way we get to work would be a humble example. Conventional census-taking, under this definition, could be characterized as a major scientific and technical paradigm.

As long as our paradigms work well for us, we tend not to change them. Occasionally, however, paradigms break down and have to be replaced. The bridge goes out and we need to find another route to work. As Kuhn pointed out in his seminal book on the structure of scientific revolutions, paradigms break down in science, as well (Kuhn 1970). Perhaps the most famous example of this is the revolution in the thinking of astronomers that occurred when the Ptolemaic earth-centered view of the universe was replaced by the Copernican view of an earth that revolved, with the other planets, around the sun.

Kish, in his paper, argues that major problems exist with the conventional census-taking paradigm. He then goes on to consider two possible alternatives: rolling censuses and administrative registers. My objective here will be to round out and occasionally balance Kish’s presentation of these topics.

Conventional Census-Taking

Conventional censuses, like those in Canada and the U.S., continue to do many things very well. Indeed, at present, we have no adequate substitute for them; nonetheless, Kish’s point of view on the need for at least some change seems compelling. Rising costs are a big factor. There have been many improvements in census-taking in this century; still, in both Canada and the U.S., total costs and even costs per person have risen significantly:

- The 1990 decennial census in the U.S. is budgeted at about \$10 (U.S.) per person. Even adjusting for inflation, this is a four-fold increase over what the per capita expenses were in 1960. Item content differences between the two censuses are small and essentially not a factor in explaining the difference. Both the 1960 and 1990 Census, for example, asked only 7 population questions of everyone (U.S. Bureau of the Census 1989). The Census long-form sample in 1960 contained 35 questions and was to be completed by 25% of the population. For 1990, the Census long-form sample was given to 16% of U.S. households and had 33 questions.

¹ Fritz Scheuren, Director, Statistics of Income Division, Internal Revenue Service. The opinions expressed here are those of the author and do not necessarily represent the position of the Internal Revenue Service.

- The situation in Canada is similar with regard to the costs of census-taking. For example, the 1991 Canadian Census is budgeted at about \$9.50 (CAN) per person. Like the U.S. Census, there are again just 7, albeit somewhat different, population items that are asked of everyone. Like the 1990 U.S. Census, questions on housing are included for everyone (2 in Canada and 7 in the U.S.). In Canada, a 20% long-form sample will be employed in 1991. The Canadian long-form questionnaire has 45 items for 1991. The 1961 census in Canada was quite different from that planned for 1991 and, thus, meaningful cost comparisons are hard to make. Nonetheless, looking back 30 years in Canada, the same long-term trend in census-taking costs seems to exist; however, per capita costs have been roughly the same – even declining slightly – in the last two or three censuses.

The U.S. Census Bureau has looked at the growing cost of conventional census-taking and concluded that a major change may be needed (Browne 1989). Labor costs have grown appreciably in recent decades in both Canada and the U.S. Technological improvements have not been great enough to offset these costs, though some, like TIGER (Topographically Integrated Geographic Encoding and Referencing) and CATI (Computer-Assisted Telephone Interviewing), offer promise. Greater attention in the U.S. to improved population coverage is another important factor (Anderson 1990). The degree of public cooperation in the census also seems to be dropping, at least as reflected by the poorer than anticipated mail response rate for the 1990 U.S. census. (It should be noted that, in Canada, public cooperation has fluctuated, with no clear tendency.)

Increasing cost is not the only major problem facing conventional census-taking. Perhaps of even greater importance, as Kish notes, is the growing rate of obsolescence of the information collected. The combination of rising costs and growing information obsolescence has had the effect of reducing the benefit/cost ratio for conventional censuses steadily and dramatically.

To obtain more frequent small area data, some countries have introduced quinquennial censuses. For example, in Canada this was first done nationally in 1956. Budget problems led to the 1986 Canadian Census being cancelled and then reinstated. Indeed, it is unclear whether there will be a Canadian Census in 1996. While a quinquennial census was also legislated in the U.S., funds were never made available.

Rolling Censuses

As Kish rightly observes, conventional census-taking, of necessity, must sacrifice both timeliness and item content (on a 100% basis) to achieve complete spatial detail and high population coverage.

One of the alternatives that Kish asks us to look at is a “rolling census.” His proposal envisions the sampling of a country over a decade in such a way that every area is eventually covered. In its purest form, space and time become a single dimension and content remains fixed, such that, at decade’s end, we have obtained cumulative information on the entire country for a given set of items.

The chief advantage of a rolling census is that it can avoid the problem of information obsolescence at national and major subnational levels. For small geographic areas, though, there would, of course, still be only one observation per decade. Unlike a conventional census, comparisons among small geographic areas would be very difficult to interpret because the data are being collected at different points in time (Fellegi 1981).

For a rolling census or survey, unit costs could be higher, as Kish notes, than in a more conventional enumeration (indeed, *ceteris paribus*, maybe even higher than the cost of existing survey efforts). In an age of fixed or declining resources, therefore, it might not be possible

to do a complete “enumeration” each decade, even if content were significantly scaled back. Rolling samples would seem to have their greatest attractiveness not as a replacement for conventional censuses, but, say, as part of a strategy to link together census-taking with ongoing surveys and local area population estimates for the intercensal years (Herriot, Bateman and McCarthy 1989).

Both the United States and Canada employ monthly surveys to estimate the national (and some subnational) labor force characteristics. The Canadian Labor Force Survey (LFS) of 64,500 households covers 0.67% of the total Canadian population each month. “Given the rotation pattern in effect for the LFS, the 0.67% sample per month rolls up into a 6.7% sample of unique households over a 5-year period” (Drew 1989). In the Canadian context, at least, Kish’s proposal may be feasible. A sample survey vehicle could be designed, with some reduction in the month-to-month household overlap, which could achieve many of the benefits he has stated for a rolling sample, while also meeting the information needs currently met by ongoing household surveys (Drew 1989). This sample would not replace the 100% census count data, itself, but, might be a *partial* substitute for Canada’s 20% long-form census sample.

Because the United States has a population about 10 times larger than Canada, the tradeoffs involving rolling samples and overall country coverage are not as favorable as they are in Canada. The U.S. Current Population Survey (CPS), for instance, at about 60,000 households, covers only .06% of the total U.S. population monthly. Even if cumulated over a *whole* decade (but, with no change in its rotation pattern), the CPS would cover just roughly 1% of all U.S. households. This does not compare well in size to the overall 16% long-form sample being conducted as part of the 1990 U.S. Census.

To bring the rolling sample population coverage nearer to the 1990 U.S. decennial sample, major changes in the CPS rotation pattern, like those Kish asks us to look at, would be needed. Other U.S. Census Bureau surveys might also have to be redesigned if the objective were to achieve even a partial substitute. Despite these changes, moreover, the resulting decade-long sample would still be only a small percent of the total U.S. population – perhaps, at best, in the 2% to 3% range, assuming resources and other requirements remained essentially fixed.

In both Canada and the U.S., the likely higher unit costs of a rolling sample may need to be addressed by changes in survey procedures: how area segments are listed (Royce and Drew, 1988); how first contact with households is made, *etc.* Where is it written, for example, that a personal interview contact is needed before using other modes of collection?

It will be no mean challenge to keep *effective* sample sizes equal for the major level and change components now obtained from ongoing surveys (*e.g.*, Tegels and Cahoon 1982). Some compromise may be needed, moreover, in the extent to which the basic content of the current long-form Census samples can be included. Despite these challenges, or perhaps because of them, Kish’s ideas on rolling samples deserve continued serious attention and should be the focus of extensive practical experimentation.

Administrative Registers

With the flowering of scientific sample survey methods in the 1940’s (Bailar, 1990), the use of administrative records for statistical purposes became relatively less important in many national statistics programs. By the early 1980’s, however, at least in the developed countries, the pendulum had begun to swing back. Kish recognizes this trend and rightly quotes Philip Redfern, who has been the major chronicler of this phenomenon internationally (Redfern 1987). While the Danes seem to have gone the farthest (Jensen 1983 and 1987), major efforts have been made in Canada (*e.g.*, Statistics Canada 1990) and even some in the U.S. (*e.g.*, Alvey and Kilss 1990).

A good summary of most of the key barriers to the greater use of administrative registers for census-taking is found in Redfern (1989), including the extensive discussion published with that paper. Perception barriers by the citizens (*e.g.*, in Germany) are mentioned as problems. Psychological barriers by the national statistical service may, however, be of equal or even greater importance. Major scientific “paradigm shifts” generally have this problem (Kuhn 1970). Certainly, this seemed to be part of the reason for the reception given to the proposal (made by me in 1980) to explore the feasibility of making administrative records an integral part of the U.S. Census of Population. While a sketch of such a proposal was eventually given at the 1982 American Statistical Association meetings (Alvey and Scheuren 1982), it seems, with a few fairly limited exceptions (*e.g.*, Irwin 1984, Citro and Cohen 1985), that serious interest at the Census Bureau has been notably lacking.

Suffice it to say that in the U.S. very little of the needed research has been undertaken. This is true, despite continuing efforts to give the proposal prominence (Jabine and Scheuren 1985 and 1987) and to get it discussed widely (Butz 1985). Sadly, therefore, I have to agree that Kish is probably right that in the United States, at least for the year 2,000, “. . . we should not expect [administrative registers] to replace censuses.”

The 1990 U.S. decennial census could have been used as a proving (or disproving) ground for some of the needed research into administrative record alternatives. Why that didn’t happen is a matter that can only be speculated about. A contributing factor, quite possibly, is a case of “paradigm paralysis” (Barker 1988). The literally decades-long controversy about whether to adjust census “counts” seems to have locked the U.S. Bureau of the Census into what some, at least, would call an increasingly sterile intellectual position (Fienberg 1990). The viewpoint that they have adopted makes it very hard for them to see any alternative, like a (partial) administrative record approach, that starts out with the notion that adjustments would be required.

The situation is different in Canada. Since the late 1970’s, Statistics Canada has assembled many of the building blocks needed to conduct an administrative record census (*e.g.*, Drew 1989; Podoluk 1987; Verma and Raby 1989). While much remains to be done, such a change could even happen as early as 1996. For example, the coverage of the Canadian tax return system, alone, is quite high and growing. In 1987, for instance, it has been estimated that the coverage was about 94% – *i.e.*, about 3% less than the 96.8% coverage achieved in the 1986 Canadian Census. By 1991, tax return coverage, alone, should be up to about 97% or better, with overall administrative record coverage still higher and likely to grow further in the 1990’s.

Kish expresses concern that administrative registers, even after they become adequate in quality and coverage, will “supply only a few, bare demographic variables: head counts, age, sex and little more.” An immediate observation concerning his remark is that conventional censuses do *little* more than this, themselves, at least for the 100% items. It is also evident that, while the variables on administrative records are not the same as those collected in a traditional census, there may *already* be more available than Kish realizes (*e.g.*, Meyer 1990; Alvey and Scheuren 1982).

More important even than any current item content comparison is the need to emphasize that the proposal to use administrative registers in census-taking does not envision that administrative records have to be used as they are. *Administrative records will need to be changed.* In my personal opinion, limited optimism about achieving needed changes is justified. However, without a doubt, it is too much to expect of administrative records that they will be able to capture exactly the same concepts now measured in censuses and surveys. Additionally, there almost certainly will need to be special efforts, using existing census-taking techniques, to separately enumerate certain groups. The efforts in the 1990 U.S. Census to count the homeless would be one such example.

Censuses and administrative records each have inherent limitations. Unavoidable conceptual differences will be a major barrier to any shift from one medium to another. Administrative feasibility is another issue; however, some hard-to-duplicate census concepts (*e.g.*, households) may not be as important to the measurement process as formerly.

Shifts in methodology (from a conventional census to administrative records) for some uses would potentially be accompanied by a parallel shift in the underlying concepts measured. Some concepts may alter or expand in meaning, including our ability to measure them (*e.g.*, families). We also must ascertain the extent to which respondents answer survey questions the same way they fill out administrative forms that may have real direct impact in their lives.

In recent years, traditional survey methodology has been enhanced by new tools from the field of cognitive psychology. These cognitive research tools could be used to understand any conceptual differences between the meaning of terms when they are used in surveys or drawn from administrative records. We may not have what we think we have anyway (Bates and DeMaio 1989). In any case, there is already an extensive body of cognitive research that can be drawn on (*e.g.*, Dippo 1987; Fienberg and Tanur 1989; Jobe and Mingay 1990).

Kish is close to the mark when he goes on to say that administrative registers “will fail to meet the demands of modern society for richer sources of statistics.” Such demands, of course, appear to be insatiable. Even if they were not, administrative records will never have the flexibility and responsiveness of surveys. Registers, however, (including partial ones like those that exist in the U.S.) when linked to survey data, can be extremely important as auxiliary variables in making improved direct national survey – and even subnational survey – estimates. The U.S. Census Bureau’s Survey of Income and Program Participation research on the use of Internal Revenue Service data for improving the precision of national survey estimates is a good recent example (Huggins and Fay 1988). Indirect (*e.g.*, synthetic) estimates for small areas would still be needed for variables not on the administrative registers (Platek, Rao, Särndal, and Singh 1987). The registers, though, might provide a source of valuable symptomatic indicators.

Concluding Observations

The case Kish makes for considering a “paradigm shift” in census-taking seems compelling, at least in developed countries like Canada and the U.S. The rolling census alternative he proposes is probably too expensive to fully implement as a complete substitute for a census. Rolling samples do offer real promise, however, if they can be integrated into the current ongoing survey operations of Canadian and U.S. national statistical programs. Such samples could provide a needed link in addressing small area estimation needs that might otherwise not be met. Less promising, but still possible, is their use as a (partial) substitute for the census long-form samples.

Kish may be unduly pessimistic about administrative registers. The Canadian situation, however, differs from the United States:

- In Canada, it is already within the realm of feasibility to combine rolling samples with administrative records as an alternative to conventional census-taking. This is not to say that enormous practical challenges don’t remain. The 100% count portion of the Canadian census, though, could be done with administrative records as a starting point, augmented by a large-scale survey to measure and potentially adjust for undercoverage. The Canadian 20% census long-form sample might be, at least partially, replaced by a rolling sample. The content of the Census long-form is considerably richer than that of household surveys, but the content differences could be made up through additional questions “piggy-backing” the on-going surveys at regular intervals. Coverage issues surrounding the use of administrative records could also be addressed directly with rolling samples, especially to calibrate for changes in administrative records between censuses.

- In the United States, the U.S. Census Bureau has begun to look at alternatives other than conventional census-taking (Bounpane 1988). Unfortunately, the research needed to look at an administrative register alternative has barely begun. Whether the Census Bureau will find a better approach than the use of administrative records and rolling samples remains to be seen (Browne 1989). Whatever other alternatives they study, however, the use of administrative registers as a partial replacement for the conventional 100% counts definitely needs to be considered. A preliminary research agenda updating earlier ideas is given in Scheuren, Alvey and Kilss 1990.

Kish is right in saying that, with the radical proposals he (and I) are discussing, the answer is uncertain. Like him, I believe that "the balance of variance components" favors a change from conventional census-taking in most cases. "However, theoretical as well as empirical investigations will be needed to decide matters."

In a change as big as the one proposed here, the "balance" that needs to be struck goes, of course, well beyond looking at variance (and bias) components. Kish recognizes this in numerous ways in his paper. One issue that needs to be emphasized more, though, is that some aspects, at least, of the paradigm shifts being considered could go to the heart of the social contract that exists between national statistical agencies and the people that those agencies have a mission to serve. For instance, in the U.S. Constitution, there is a requirement that an "enumeration" of the population take place every ten years. Would the use of administrative records or rolling censuses fit within this "Constitutional paradigm?" Perhaps the starting place is to adopt a broader definition of "enumeration."

Another example where social contract issues arise is the extent to which the greater use of existing (or expanded) administrative data for statistical purposes might be seen as an unwelcome increase in the intrusiveness of the State into the private lives of its citizens (Grace, 1989). As legitimate as concerns about "intrusiveness" might be, though, there is no evidence in a North American context, at least, that they pose an insurmountable barrier. On the contrary, there have been virtually no adverse public reactions to past U.S. additions to administrative records for statistical purposes (*e.g.*, of residential address information in 1972, 1974 and 1980 tax returns). To my knowledge the issue, so far, has not come up directly yet in Canada, at least at the Federal level.

In summary, to make changes of the types being discussed by Kish, there is, as he points out, the need for a lot more scientific research. Studying the implementation technologies will be an even bigger job. Finally, the issues go beyond our profession and may well be settled in other arenas. Wherever they are decided, it is incumbent on us, as statisticians, to frame the debate in terms of feasible options. Kish has taken us a long way down that path and is to be greatly congratulated.

REFERENCES

- ALVEY, W. and KILSS, B. (eds.) (1990). *Statistics of Income and Related Administrative Record Research*. U.S. Department of the Treasury, Internal Revenue Service. See also Kilss, Beth and Alvey, Wendy (eds.) (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, vols. 1 and 2, U.S. Department of the Treasury, Internal Revenue Service.
- ALVEY, W. and SCHEUREN, F. (1982). Background for an Administrative Record Census. *Proceedings of the Section on Social Statistics, American Statistical Association*, 137-146.
- ANDERSON, M. (1990). 'According to their respective numbers . . .' for the twenty-first time. *Chance*, 3, 12-18.

- BAILAR, B. (1990). Contributions to Statistical Methodology from the Federal Government. *Survey Methodology*, 16.
- BARKER, J.A. (1988). *Discovering the Future: The Business of Paradigms*.
- BATES, N.A., and DEMAIO, T.A. (1989). Using Cognitive Research Methods to Improve the Design of the Decennial Census Form. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 267-285.
- BOUNPANE, P. (1988). A Sample Census: A valid alternative to a complete count census? 46th Session of the International Statistical Institute.
- BROWNE, D.L. (1989). U.S. Bureau of the Census: Facing the future labor shortage. *Asian and Pacific Population Forum*, 3, 4.
- BUTZ, W. (1985). Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities. *Journal of Business and Economic Statistics*, 393-395.
- CITRO, C., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. National Academy Press, Washington, DC.
- DIPPO, C. (1987). A Review of Statistical Research at the U.S. Bureau of Labor Statistics. *Journal of Official Statistics*, 3, 289-297.
- DREW, J. D. (1989). Address Register Development and its possible future role in Integration of Census, Survey and Administrative Data. A paper presented at the U.S. Bureau of the Census/Statistics Canada Interchange. (Unpublished).
- FELLEGI, I.P. (1981). Discussion of a paper by Leslie Kish entitled "Population Counts from Cumulated Samples." *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau. An Analysis, Review and Response*, Congressional Research Service, the Library of Congress.
- FIENBERG, S. (1990). An Adjusted Census in 1990? An Interim Report. *Chance*, 3, 19-21.
- FIENBERG, S., and TANUR J. (1989). Combining Cognitive and Statistical Approaches to Survey Design. *Science*, 243, 1017-1022.
- GRACE, J.W. (1989). The Use of Administrative Records for Social Research. Statistics Canada Workshop, December 12, 1989, Ottawa, Ontario.
- HAMMOND, R.B. (1990). The 1990 Decennial Census: An Overview. *Conference Proceedings, Advanced Computing for the Social Sciences*, sponsored by the Oak Ridge National Laboratory and the U.S. Bureau of the Census, April 10-12, 1990, Williamsburg, Virginia.
- HERRIOT, R., BATEMAN, D.V., and MCCARTHY, W. F. (1989). The Decade Census Program – New Approach for Meeting the Nation's Needs for Sub-National Data. To appear in *American Statistical Association Proceedings, Social Statistics Section*.
- HUGGINS, V., and FAY, R. (1988). Use of Administrative Data in SIPP Longitudinal Estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.
- IRWIN, R. (1984). Feasibility of an Administrative Records Census in 1990. Special report on the use of administrative records, committee on the use of administrative records in the 1990 Census, unpublished Census Bureau report.
- JABINE, T.B., and SCHEUREN, F. (1985). Goals for Statistical Uses of Administrative Records: The Next Ten Years. *Journal of Business and Economic Statistics*, 380-391.
- JABINE, T.B. and SCHEUREN, F. (1987). Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going? *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, 43-72.
- JENSEN, P. (1983). Towards a Register-Based Statistical System – Some Danish Experiences. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.

- JENSEN, P. (1987). The Quality of Administrative Data from a Statistical Point of View: Some Danish Experience and Consideration. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs and M.P. Singh (eds.) Statistics Canada, Ottawa.
- JOBE, J.B., and MINGAY, D.J. (1990). Cognition and Survey Measurement: History and Overview. *Applied Cognitive Psychology*, in press.
- KUHN, T.S. (1970). *The Structure of Scientific Revolutions*. Second Edition, Enlarged, The University of Chicago Press, Chicago.
- MEYER, B. (1990). The Tax System: Comparisons of Demographic, Labour Force and Income Results for Individuals and Families. Small Area and Administrative Data Division, Statistics Canada.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, E.E., and SINGH, M.P. (1987). *Small Area Statistics*, New York: Wiley-Interscience.
- PODOLUK, J. (1987). Administrative Data as Alternative Sources to Census Data. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*; J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, 273-290.
- REDFERN, P. (1987). A Study of the Future of the Census of Population: Alternative Approaches Eurostat Theme 3 Series C. Luxembourg: Office for Official Publications of the European Communities.
- REDFERN, P. (1989). Population Registers: Some Administrative and Statistical Pros and Cons. *The Journal of the Royal Statistical Society, Series A* (Statistics in Society), 152, 1-41.
- ROYCE, D., and DREW, J.D. (1988). Address Register Research: Current Status and Future Plans. 1991 Research and Testing Project, 1991 Census, Statistics Canada, Ottawa.
- SCHEUREN, F., ALVEY, W., and KILSS, B. (1990). Paradigm Shifts: Administrative Records and Census-Taking.
- STATISTICS CANADA (1990). Research papers and reports. Bibliography, Small area and administrative data division, Ottawa, Ontario. (unpublished).
- TEGELS, R., and CAHOON, L. S. (1982). The Redesign of the Current Population Survey: The investigation into alternate rotation plans. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- U.S. BUREAU OF THE CENSUS (1989). *200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC.
- VERMA, R.B.P., and RABY, R. (1989). The Use of Administrative Records for Estimating Population in Canada. *Survey Methodology*, 15, 261-270.

Comments on Articles in the Special Section

MORRIS H. HANSEN¹

These are excellent papers that I enjoyed reading. Three of these papers focus primarily on historical and current developments and to some extent looking to the future. The paper by Kish is focused on and is an effort to influence some important future developments. I will attempt to add a little clarification from my own personal history and point of view on the historical summaries, and a little perspective, again from my personal point of view, on Kish's proposal for rolling censuses to replace the more traditional censuses.

Rao and Bellhouse have given a compact but useful survey of sampling development. Their summary begins, after a few preliminaries, at about the time that I first began to participate in censuses and sample surveys, and their improvement.

Their survey is done about as well as can be accomplished in such a compact summary, without elaborating on details. However, I would like to provide a slightly different view than they present on the development of sampling with probabilities proportionate to size or to measures of size (PPS). They accurately indicate that we (Hansen and Hurwitz) developed the theory for PPS sampling with replacement as an approximation. We were unsuccessful in solving the problem of variance estimation with varying probabilities when sampling without replacement that was soon solved by Horvitz-Thompson and others. However, with possibly rare exceptions, we never proposed the use of or used sampling with replacement. In practice, we did PPS sampling without replacement, usually either by choosing two or more units from a stratum by a systematic sampling procedure with the units arranged in a random or systematic sequence, or by choosing one unit per stratum. Units that would have had high probabilities of selection were selected with certainty. We prepared estimates of aggregates and functions of these by weighting by the reciprocals of the probabilities, exactly as in what has come to be referred to as the Horvitz-Thompson estimator. The variance estimators resulted in moderate overestimates because they assumed sampling with replacement as a simplification. Ordinarily, we have not regarded moderate overestimates of variance as a serious concern. The ultimate cluster variance estimator was often used. This is a very simple approximate variance estimator that involves weighting (if subsampling has been used) within the first stage units up to the first stage unit level, and then computing the variance between such first-stage unit estimates (see Hansen, Hurwitz, and Madow, p. 257). Horvitz and Thompson provided the initial breakthrough in variance estimation when sampling more than one unit per stratum with varying probabilities.

Sampling with PPS had the advantages that Rao and Bellhouse briefly describe. In addition, its use was a great convenience in multistage sampling, with probabilities proportionate to measures of size at each stage up to the final. The probabilities at the final stage were often set to achieve uniform overall probabilities of selection of the elementary units.

I add one other comment on their paper with respect to jackknife variance estimation. They indicate that the jackknife variance estimators are known to be inconsistent for nonsmooth functions like quantiles, even in the case of simple random sampling. They might have said, especially in the case of simple random sampling of the elements that are the units of analysis. We have recently demonstrated empirically that variances of medians and (in this case)

¹ Morris H. Hansen, Westat, 1650 Research Boulevard, Rockville, MD, 20850, U.S.A.

of 10th and 90th percentiles can be well estimated with the usual ultimate cluster jackknife variance estimation procedure with multistage sampling in which two or more first-stage units or combinations of them are identified in a stratum (one dropped and the other doubled, to form a replicate). We hypothesize that jackknife worked well in these applications because each ultimate cluster associated with a first-stage unit contains a substantial number of elementary units in the sample. We anticipate that it would work equally well, although we have not demonstrated it, when the jackknife replicates are formed by another procedure often followed, in which a simple random (or stratified random) sample is divided into m simple random subsamples (or stratified random subsamples utilizing the same strata to the extent feasible), and dropping one subsample at a time.

Fienberg and Tanur have presented an interesting perspective on the influence of the institutional setting in which survey research has developed. I agree with their view that an improved understanding of the development of survey methods is achieved by an understanding of the institutions through which survey research and surveys are done. At least those survey developments in which I have participated have arisen largely out of the institutional setting, and the need and opportunity to solve problems that occurred in accomplishing programs of the institution. Again, I have comments on some of the details in the developments in which I was a participant.

Fienberg and Tanur properly indicate that the design of what is now known as the Current Population Survey or CPS (earlier known as the Labour Force Survey) had a key role in the evolution of sampling theory and its application that has influenced other developments. However, they incorrectly suggest that its principal origins were in the experimental Trial Census of Unemployment carried out in late 1933 and early 1934 as a Civil Works Administration (CWA) project in three cities. There is some confusion in their paper of the 1933-34 CWA trial census with the 1937 "Enumerative Check Census" that accompanied the 1937 "Unemployment Census". It was the latter that, as they mention, Dedrick, Hansen, Stouffer, and Stephan jointly worked on, and that was the progenitor of the CPS. The 1937 Unemployment Census was a national registration done through the Post Office. The Enumerative Check Census was taken by mail carriers in a national probability sample of postal routes – they took a complete census of each postal route in the sample. New concepts for measuring labour force and unemployment were developed and applied in it based on behavior in a prior week. It was also a first application of nationwide area probability sampling. Its purpose was to evaluate the 1937 national registration of the unemployed (as discussed in the accompanying paper by Barbara Bailer). That sample survey taught us much, and was the seed for the monthly Labor Force Survey, later to become the Current Population Survey. Again, I was an active participant. Bailer describes it well. Stock, Frankel, and Webb and others at the Work Projects Administration (WPA) also had a role in the design of the national registration and of the Enumerative Check Census. Those were the days of dire unemployment, and the need for a continuing measure was obvious and urgent.

With this experience Stock, Frankel, and Webb, along with their colleagues at WPA perceived the opportunity and need for a continuing survey. They initiated a monthly unemployment and labor force survey, introducing some imaginative concepts in survey design (but also some problems that needed later correction). The monthly survey was just getting well established when Pearl Harbor and U.S. entry into World War II occurred, and the needs for information were radically changed. Labor shortage rather than high unemployment became the problem. The WPA was no longer needed and was abolished, and the survey was transferred to the Bureau of the Census to become a labor force survey to measure especially war-time implications of labor force participation and employment. When the survey was transferred to the Bureau of the Census we perceived some problems in the original design and developed

solutions to them, which led to the introduction, among other things, of PPS sampling and other design innovations. These developments for the labor force survey (now the CPS with a much broader role) have had a substantial impact on sample methodology, and more important, on meeting the needs of the nation for up-to-date information, not only on labor force but on many other subjects – demographic, social, and economic.

Feinberg and Tanur might also have emphasized the remarkable consequences of bringing together census-taking and sampling, along with computerization and automated reading of position marks on census questionnaires. In modern censuses in the United States, beginning with the 1960 Census, the questionnaires used for collecting information from all households are relatively brief in content. The principal content of the censuses is now obtained through samples taken simultaneously with and as part of the census, and, of course, on an exceedingly large scale in order to produce useful data for perhaps 40,000 small areas. A related development was the introduction in the 1960 Census of self-enumeration methods. The decision to introduce self-enumeration was guided by the application of the response error model to which Feinberg and Tanur refer, and by associated research and experiments on response errors, and especially on the correlated response errors associated with the work of enumerators. These innovations were guided by large-scale experiments that were done prior to and as part of the 1950 Census and in later censuses as well as in separate experiments. Another contribution was FOSDIC (Film Optical Sensing Device for Input to Computers), a device for reading position marks designed by the Bureau of Standards at the Census Bureau's request, in response to Census Bureau needs to replace the massive key-punching effort in a census. A consequence of the innovations that were introduced was more timely results and generally more accurate censuses, as well as lower costs. The opportunities for progress arose in view of the problems of large-scale census taking, and how they might be solved with the application of sampling and self-enumeration, along with the remarkable advances made possible by the development and application of electronic computers and FOSDIC, in which the Census Bureau was a pioneer.

In the late 1930's, some of the top Census Bureau staff, as well as members of Congress, were reluctant to see sampling introduced into the work of the Census Bureau. Complete enumeration had been the tradition. The use of probability sampling in the 1937 enumerative check census associated with the national unemployment registration was an important factor in achieving the acceptance of sampling as a methodology appropriate to the Bureau of the Census, again as more fully told in the accompanying paper by Bailar. The 1940 population census was a pioneering effort in the application of sampling in the collection of supplemental items of information in a census. In this effort Deming and I worked as colleagues. I was working with Calvert Dedrick, and Deming with Philip Hauser, with effective consultation and advice from Fred Stephan, and we all worked as a team in developing this important milestone in the application of sampling.

I have little in the way of comments to add to the paper by Barbara Bailar. As the paper indicates, I was an active participant along with Bill Hurwitz and our colleagues, in the developments she describes so well. I do have a minor correction. Feinberg and Tanur correctly identify the 1951 paper on response error models by Hansen, Hurwitz, Marks, and Mauldin as the original publication on the model, which Bailar credits to a later (1960) paper by Hansen, Hurwitz, and Bershad. The later paper elaborated those results, and included empirical data from the application of the model in large-scale randomization experiments involving the random assignment of enumerators in the 1950 Census. Analysis of these results as summarized in the 1960 paper showed the substantial and striking impact on small area census statistics of correlated errors within the work of interviewers. Earlier memoranda containing the results reported in that paper, and associated studies, were the principal vehicles that led

to the use of self-enumeration as the procedure for collecting the principal content items in the 1960 Census. They also led to transferring the collection of much of the information to a large sample instead of a complete census, with substantial cost reduction implications, improved timing, and generally improved quality. Bailer's paper provides an excellent summary description.

I should note, in this connection, the remarkable contribution to these developments that came from Bill (William N.) Hurwitz. He and I worked as a team that was far more effective than the sum of our individual contributions. In addition, I cannot give enough credit to our colleagues that we recruited and helped to stimulate and to some extent train, and who became the backbone of developments in the Census Bureau in the application of sampling, quality control, and operational research methods to the successful design and conduct of samples and censuses in wide ranging subject areas. Leaders among these colleagues included Max Bershad, Joseph Daly, Leon Gilford, William Madow, Eli Marks, Harold Nisselson, Jack Ogus, Leon Pritzker, Joseph Steinberg, Benjamin Tepping, Joe Waksberg, Ralph Woodruff, and others. I often get much of the credit, but without Bill Hurwitz, especially, and our colleagues, it could not have occurred.

I should mention that we benefited greatly, also, from the participation and advice from a panel of statistical consultants, with Bill Cochran (William G. Cochran) as chairman, over the years from 1955 until I left the Bureau in 1968. Other principal members included Fred Stephan (Frederick F. Stephan) and Bill Madow (William G. Madow) for the full time period, and Ivan Fellegi from Statistics Canada, H.O. Hartley, and others for part of the time. All were exceedingly able. However, we did not look to them as experts whose advice would simply be sought and generally followed. Instead, we operated on an interactive basis. We discussed specific issues or problems as well as all phases of total survey design for a particular survey, experiment or census. We received much useful advice; they also learned from us.

The paper by Leslie Kish moves the emphasis from historical background and recent and current advances to proposals for taking censuses of the future, through the introduction of what he calls rolling censuses. He also describes rolling samples in various forms.

Each of the kinds of rolling samples that he discusses, with and without overlapping panels are, as he indicates, in use for various purposes at the present time, and his discussion of these does not propose anything new. I suppose he introduces them for generality and as a means of suggesting their potential relationship to a rolling census.

The particular rolling census he describes is a weekly sample, with the total population of housing units at each point of time subdivided into 520 subsamples, one to be covered each week over a 10-year period. Thus, the entire population of housing units would be covered in a decade except for new additions of housing units in samples that had already been covered earlier in the decade. If the procedure were continued over time, then at any point in time the aggregate of the 520 samples for the prior ten years would provide average census results, representing the average situation over the prior 10-year period. It is an interesting and imaginative proposal. However, there are also problems.

He suggests a rolling census without any overlap in the coverage in successive weeks or other periods, except after the full decade when it starts all over again. Such an approach would provide a large national cross section sample each week, as well as average or aggregate results for each month, each year, and for other periods. However, without any overlap in the samples, it will be a relatively crude instrument for measuring changes occurring in small areas from week to week, from month to month, or even from year to year. Overlapping samples might be introduced, as he indicates, but would add greatly to costs. Of course, changes can be measured with the proposed rolling samples, but without partially overlapping samples the result would be large sampling errors of estimates of change for small areas. Providing data

for small areas is a primary purpose of the Decennial Census. I believe that reliably measuring such changes may be as important as providing aggregate measures for points in time. While Kish recognizes this, he seems to dismiss it.

Undercoverage of the population would likely be a particularly serious problem with a rolling census. Because of the general recognition by the public of the need for censuses, along with the intense publicity that is feasible for a census, the completeness of coverage of the censuses has traditionally been much greater than that in even the best sample surveys (although coverage still remains a problem in the censuses). The problem of net undercoverage in sample surveys is quite general – even including the Current Population Survey in the U.S. which is often taken as a model. Public interest with continuing weekly publicity for a rolling census could not conceivably be maintained.

Another issue in my judgment is the likely high cost of such a system. Kish recognizes this, also, and then seems to dismiss it. While I have not seen any cost estimates, I would not be surprised that over a decade the rolling census would cost substantially more than the cost of taking complete censuses quinquennially, plus the cost of relatively large-scale monthly samples to provide measures of change and information on various subjects for states and large areas within most states. Moreover, I anticipate that quinquennial censuses would be easier to interpret and more useful by providing measures for small areas at points in time, or for short intervals of time, rather than providing average measures over periods up to ten years.

The Census Bureau, influenced, in part, by Kish's earlier recommendations for such a rolling census, and the desire to spread the workloads has come up with some proposed alternatives for consideration for taking a brief decennial census along with rotating censuses. They consider some alternative approaches to rotating censuses of whole states over a decade. It is an innovative proposal intended to spread the workload while avoiding the high cost of a rolling census such as described by Kish.

I am one who believes that a quinquennial census, along with ongoing large-scale current surveys, are well worth a substantial cost. However, I believe that if a rolling census were adopted, as proposed by Kish, overlapping samples should be used. A rolling census, even without overlapping samples, may cost considerably more than the cost of the current census program extended to include a quinquennial census. I question if it is worth the added cost, or that it has advantages over a quinquennial census plus substantial intercensal samples. I anticipate that the rolling census approach would yield less useful information than quinquennial censuses for most purposes because it would provide complete census counts only for averages over a 10-year period. Quinquennial censuses, along with sufficiently large current samples to provide relatively up-to-date information for large areas, along with other procedures for providing data for state, county, and perhaps also small area population estimates, seem to have advantages from a cost-benefit point of view.

Kish is to be commended for his efforts to solve some of the census problems by a radical new approach. However, to me, the rolling census does not appear to be the answer. Perhaps more effective utilization of administrative records can provide results that hold more promise, again along with current samples and a decennial, or, hopefully, quinquennial censuses. Perhaps the remarkable new computerized mapping and coding system (known as TIGER) developed by the Census Bureau for the 1990 Census holds much promise for improving census-taking, and for current sample surveys. In addition, incorporating the TIGER geographic coding into the major administration records systems might make them more accessible for population estimates and for other uses. Up-to-date maintenance of TIGER, along with a currently maintained address register, are hopefully to be included in the Census Bureau's future plans.

REFERENCES

- HANSEN, M., HURWITZ, W., and MADOW, W. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.
- HANSEN, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2, 180-190.
- OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 162-179.
- DUNCAN, J.W., AND SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. U.S. Government Printing Office, Washington, D.C.

Reply

J.N.K. RAO and D.R. BELLHOUSE

We thank the discussants, Hansen and Smith, for their useful comments.

Hansen provided important observations on the development of PPS sampling. He is correct in saying that Hansen and Hurwitz (1943) did not propose the use of sampling with replacement and that only for variance estimation they assumed sampling with replacement. Incidentally, Murthy (1967, p. 184) notes that Mahalanobis (1938) has referred to PPS sampling and the associated unbiased estimator of a total in the context of sampling plots for a crop survey.

Hansen also made some interesting observations on the use of delete-1 cluster jackknife variance estimator for nonsmooth functions like quantiles. It is now well-known that the delete-1 jackknife variance estimator of a quantile is inconsistent under simple random sampling. Empirical results in Kovar, Rao and Wu (1988) indicate that it is also inconsistent under stratified simple random sampling. It is also likely inconsistent under stratified cluster sampling if the subsamples from the clusters are small or if the intra-cluster correlations are significant. In Hansen's application the subsamples from the clusters are quite large and the intra-cluster correlations very small. In this case, the delete-1 cluster jackknife variance estimator may be well-behaved in view of Shao and Wu's (1989) result that the delete- d jackknife variance estimator, under simple random sampling, is consistent, provided $n^{1/2}/d \rightarrow 0$ and $n-d \rightarrow \infty$ as the sample size $n \rightarrow \infty$.

The method of dividing a simple random sample into m subsamples, each of size d say, and dropping one subsample at a time, as suggested by Hansen, is similar to Shao and Wu's delete- d jackknife except that they consider all $\binom{n}{d}$ subsamples in constructing the variance estimator. However, the delete- d jackknife variance estimator is likely to be more stable. Shao and Wu also consider balanced subsampling requiring only b subsets of size $n-d$, where $b (\geq n)$ is the number of blocks in a balanced incomplete block design.

Smith provided some important observations on the foundational aspects of sample survey theory, in particular, on the importance of Ericson's (1969) work on Bayesian estimation of a total under exchangeable priors. In this connection, we note that equivalent results for the posterior mean and the posterior variance, under simple random sampling, were also obtained by Hartley and Rao (1968). A. Scott pointed out the similarity of the two approaches in his discussion of Ericson's paper. However, an advantage of the Hartley-Rao approach is that the inferences depend on the sample design, unlike Ericson's approach. Their approach also yields useful classical inferences. Rao and Ghangurde (1972) extended the Hartley-Rao results to stratified random sampling, double sampling with unknown strata sizes, the Hansen-Hurwitz method for handling nonresponse, and two-stage random sampling.

The GUT approach for inference, proposed by Smith looks very promising. We agree with Smith that the point estimators using the different approaches rarely differ very much in practice, and that the issue essentially reduces to the choice of a measure of uncertainty, as noted in our paper.

We also agree with Smith on the importance of measuring total survey error from ongoing surveys.

ADDITIONAL REFERENCES

- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society, Indian Statistical Institute.
- MAHALANOBIS, P.C. (1938). *Statistical report on the experimental crop census, 1937*. Indian Central Jute Committee.
- RAO, J.N.K., and GHANGURDE, P.D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association*, 67, 439-443.
- SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimations. *Annals of Statistics*, 17, 1176-1197.

Reply

STEPHEN E. FIENBERG and JUDITH M. TANUR

We are grateful to Bob Groves and Morris Hansen for their insightful comments and to the editor of *Survey Methodology* for the opportunity to update our thinking in 1990 rather than waiting for 2040. Groves and Hansen make several important points; we shall attempt to react to them in turn.

We very much like Groves' summary to the effect that governments emphasizing service for the welfare of the populace demand more information about their services than do those pursuing other goals. Consistent with this thesis is the fact that the most substantial new national survey launched in the United States during the 1980s, a decade not noted for an emphasis by the federal government on expanding welfare services, was the Survey of Income and Program Participation, one of whose primary purposes has been to monitor the impact of government welfare programs on income and assets. Moreover, as the countries of Eastern Europe democratize and turn to the West for assistance in upgrading their statistical systems, including the development of infrastructures for the conduct of large scale surveys, we see additional support for such a thesis. Thus it seems to us that Groves shares our belief that the institutional bases for survey research shape the content and direction of such surveys. Whether they provided homes or incubators for the best and the brightest seems to us akin to the nature/nurture debate – more a framework for discussion than an either-or choice. Indeed, we agree with Groves that the purposes of the various sectors shaped their choice of tasks, at least in part. In line with his urging of a cross national perspective, however, we note that institutional roles differ across countries. For example, there has been a widely-held view in the United States that the Federal government should not be in the business of collecting survey data on subjective phenomena (e.g., see Turner and Martin 1984, 31-39) – a quite different stance has been taken by the British government, especially in connection with its annual report, *Social Trends* (Turner and Martin 1984, p.4).

Groves suggests that the membranes between sectors (academic, commercial, and governmental) are less permeable than we suggest. Neither we nor he have collected systematic empirical evidence on this question, but we point again to our concept of bridging institutions which bring together representatives of the various sectors, for the interchange of ideas if not personnel. And we hasten to point out that Groves' own recent appointment to the position of Associate Director of the U.S. Bureau of the Census, as well as Hansen's movement from that position into the commercial domain back in 1968, indicate the value, if not the ease, of membrane crossing.

Groves indirectly speculates that we choose to focus on technological advances, longitudinal surveys, and cognitive aspects of surveys because these are our areas of interest and experience, and he suggests several other developments that are worthy of consideration. Of course he is correct in suggesting that we have focussed on the developments that fit with our interests, but surely technological advances as a topic subsumes Groves' first two additional areas of importance: (1) development of generalized statistical software packages and (2) existence of survey data archives. We wonder, however, if the technological advances we both note, coupled with the ubiquity of surveys that we also both note, do not have negative as well as positive consequences. For example, the complex analyses of survey data by undergraduates (or indeed

any beginners) using statistical software packages often show neither an understanding of the data being analyzed nor the appropriateness of the packaged statistical methods used.

The ubiquity of surveys is a consequence not only of the demand for information but also of the relative ease with which surveys can be carried out and the data analyzed given current technology. (And we believe that the availability of survey data for reanalysis will only increase with the advent and adoption of new storage technologies such as CD-ROM and optical disks). Such ease is a mixed blessing. As Groves notes, the 1980s have seen a growing problem of nonresponse in the United States, a pattern that manifested itself earlier and (so far) more seriously in Europe. We do not need to postulate a growing trend toward demands for privacy to explain this decline in response rates, though such a trend may well exist. We need only look at the major nonresponse problems currently being encountered in the conduct of the U.S. 1990 decennial census, in both the mail-out-mail-back and in the door-to-door phases, to see evidence to support the contention that respondents are merely getting tired of being surveyed so frequently.

Further, as Groves points out, survey research has not been central to the self-image of academe, because survey research has not fully evolved into a separate identifiable discipline, with specified standards and training criteria. Since there are no departments of survey research on university campuses, almost anyone who cares can mount a survey or carry out analyses of survey data. While some people do these tasks well, others do them poorly thereby giving the whole survey enterprise a bad name. Thus, if we are to present the optimistic report on the state of the survey enterprise in 2040 that Groves envisages, it seems to us that the innovations in education and training that neither he nor we are currently able to chronicle will have to become institutionalized.

We are especially pleased to have Hansen's embellishment on our brief account of the development of the survey enterprise in the U.S. government in the 1930s and 1940s. His comments supply some of the human drama that Groves says is lacking in our institutional focus.

Hansen also expands on our account of the link between censuses and sampling and the introduction of self-enumeration into U.S. censustaking, that was guided by the study of response errors. The major decline in completion rates for self-enumeration in the 1990 decennial census suggests the need to reexamine the implications of the various components in the Hansen-Hurwitz-Marks-Mauldin model for non-sampling errors. In addition we note that as part of the 1990 census, the Bureau of the Census will mount a new Post-Enumeration Survey (PES) of 150,000 households whose results will be used to evaluate census coverage. The technological advances in computerized data management and in computer-based matching of files between the PES and the census were essential ingredients to the launching of this major new government survey and its planned use to measure both under-and over-coverage of the household-based population.

ADDITIONAL REFERENCES

- TURNER, C.F., and MARTIN, E. (eds.) (1984). *Surveying Subjective Phenomena. Vol. 1*. New York: Russell Sage Foundation.

Reply

BARBARA BAILAR

The comments from the discussants describe even more contributions of the Federal Government to the world of statistics. I am very grateful to Gordon Brackstone and Morris Hansen for mentioning these additional topics. The topic I omitted that may have had the biggest impact on statistics as well as other quantitative fields was the development of the computer for data processing and data analysis purposes. Again, the team of Hansen and Hurwitz were the prime movers, urging and funding the development of UNIVAC I and then bringing it into the Census.

Morris Hansen describes the remarkable team at Census who worked with him and Bill Hurwitz on so many topics. I feel very fortunate that I began my career at the Census Bureau when these people were there and that I was able to work with most of them for many years. It is rare that one gets that kind of apprenticeship.

Gordon Brackstone questions whether the statistical methodology developed by the Census Bureau had a benefit to the wider world of statistics. Certainly, given the amount of interaction among government statistical offices, the Bureau of the Census has influenced government statistical operations in other countries. Brackstone finds the impact of the Census Bureau development on university statistics departments rather mixed. He may be correct as far as course offerings are concerned, but I believe the ASA-NSF-Census Fellowship program and the Agriculture Fellowship program have had a big impact. More university professors and graduate students are aware of and working on non-sampling error, disclosure avoidance, and time series problems. The recent addition of Fellowship programs at the Bureau of Labour Statistics and the National Center for Education Statistics have also highlighted these research areas. The NSF now receives many proposals based on research started at one of the government agencies.

The main problem now is to make sure that research results are used. Many government programs are slow to accept new methodology because change is disruptive. Yet, to make sure that methods are improving, change is necessary.

Reply

LESLIE KISH

In his fine discussion Fritz Scheuren complements our comparisons of alternative census methods by advocating administrative registers for the USA. I support his expert plea to study what these methods could offer as additions, as complements to the decennial censuses. They are coming to many countries and we would like to know where, when, and how? It is even likely that they will not only complement, but even replace decennial censuses soon in some places. When in the USA? I don't know; we were comparatively slow and late in adopting a successful registry of births and deaths. And even now their reporting is rather slow.

Rolling samples could be designed for quick reporting, and timeliness is only one of the advantages of rolling samples. Thus it is biased to compare rolling censuses with traditional censuses, both as regards costs and benefits, only on the basis of the single output for which decennial censuses are designed. It would take detailed, technical investigations to compare the factors of costs, coverage, timeliness, content, *etc.* of rolling versus decennial censuses in the USA. But 10 to 15 million dollars monthly can go far. The issue of adequate censuses is most salient in 1990 in the USA and elsewhere, but the other uses of samples should not be forgotten, as we plan for the last decade of our twentieth century.

My contribution aims mainly to advance the *diverse* advantages of cumulations from periodic samples, which have been neglected in favor of the other benefits that can be obtained from the growing numbers of periodic surveys. Rolling censuses may become someday one of those benefits, and rolling samples have been used already – though not often enough, I believe. Asymmetrical cumulations may exist rarely and obscurely, and the split-panel designs that I propose, not at all.

Furthermore my scope is not merely national (the USA), nor even continental (North America): it is intercontinental and international. For example, registers have come to the Nordic countries and they may come to Canada before the USA. Rolling censuses pose a much smaller expansion of the Labour Force survey in Canada because it is one-tenth the size of the USA, as Fritz and I both show. But some other country may well use them before either.

Not only international, rolling samples and cumulations are also aimed to be interdisciplinary, not only for making population counts. Good many of the other needs of statistical offices – and *of other institutions for data collections!* – would be better served by a trained “permanent” staff than by a hurriedly hired huge army whose training time roughly equals their brief employment.

Scheuren is most complimentary when he calls rolling censuses a new paradigm. It is true that, as all new paradigms, they meet three big mental blocks when I present cumulations and rolling samples: a) averaging of variable data instead of an arbitrary date like April 1, of the decennial year; b) accepting some of the mobility of human populations instead of fixing them to unique sites; c) rolling samples to replace fixed primary sampling areas. So it may seem paradoxical when Morris Hansen notes that my “discussion of these does not propose anything new.” Hansen may have encountered all of these proposals, and perhaps dismissed some of them. Personally I have described rolling samples since at least 1961 and proposed rolling censuses since 1965. But I also found that for many people they come as new ideas, and often as strange new ideas.

Finally let me only add two important origins in the '40's for sampling, although for me personalities and priorities are only minor aspects of the history of any science. Iowa State at Ames should be mentioned, where, under George Snedecor and Henry Wallace, Bill Cochran started in the spring of 1939 the first course of sampling and turned out pioneer MA's, then PhD's in sampling. Then Henry Wallace (again) in the US Dept. of Agriculture started the Division of Program Surveys, hired me in 1941 and Steve Stock in 1942 for the first national samples in Washington in 1942, followed by the 1943 sample at the USBC. Stock, Frankel and Webb (from the WPA samples) began the second sampling course in fall 1939 at the USDA graduate School, which became famous and productive under Hansen, Hurwitz and their Census staff. Among influential courses there I shall testify especially to those of Deming, the major figure at the school. The teaching and learning of samples in the forties was done mostly at Ames and in the USDA, as well as at the USBC.

Some Developments of Sampling Techniques and their Use in Official Statistics in Sweden

TORE DALENIUS and CARL-ERIK SÄRNDAL¹

In this paper we present some important features of the history of sample surveys in Sweden, and we comment on related developments of sampling techniques (methods and theory) in official statistics. The account is organized into three periods as follows: (i) before 1900; (ii) 1900-1950; and (iii) after 1950. The emphasis is on the third period.

I. THE PERIOD BEFORE 1900

1. A summary view. As described in Dalenius (1957), there was a noticeable resistance against sample surveys in traditional fields of official statistics, especially among statisticians in leading positions. Sample surveys were considered justified primarily in cases where circumstances did not admit *total* surveys. In other fields there were, however, signs of appreciation, as illustrated in the next section.
2. Two classic illustrations. In the 1820's, the area of meadowland in Sweden was estimated using the following technique. For each county separately, the ratio of meadow acreage to arable land was computed for a sample of farms. This ratio was then applied to the total arable land acreage of the county, for which a separate estimate was available. And in 1830, the proposal was made by an official in a forestry board to estimate the volume of timber in a forest by means of a "strip survey method".

II. THE PERIOD 1900-1950

3. The main features. The potential of sample surveys in official statistics was slowly being understood. To the extent that sample surveys were used during this period, the design typically called for systematic sampling, whenever this was operationally feasible. In many applications, the sampling fraction was 1/10 or 1/5. In the 1940's, a major factor favouring total surveys was the war-time economy with its regulations and rationing. This influence, which lasted roughly until the end of that decade, was however counteracted by the introduction of Gallup polls into Sweden and especially by the spectacular accuracy of the Gallup Institute's forecast of the 1944 election. In particular, these trends were followed with interest by official statisticians.
4. The 1911 Forest Survey in Värmland. The essential feature of the design was that the volume of timber was measured on sample plots along 10 meter wide strips covering the area of Värmland. It is worth noting that the "representative characteristics" of the survey were analysed by means of probability theory.

¹ Tore Dalenius, Brown University, Carl-Erik Särndal, Université de Montréal.

The circumstances did not permit the authors to discuss the contents of this paper with representatives of Statistics Sweden.

5. The 1911 Housing Survey in Göteborg. This survey was carried out by the municipal statistical office in Göteborg. The selection of the sample of apartments was based on an urn scheme. Each building in Göteborg was represented by a slip with identification data. The slips were thoroughly mixed in an urn and a 20% sample of slips was selected. The motive behind the scheme was to avoid that the survey be criticized for using a biased sample. The urn scheme was described by the person in charge of the survey as the only method "which can be called representative".
6. The 1935-36 Partial Population Census. This sample census used an elaborate scheme of controlled selection. The results from this census played a decisive role in an intense debate in Sweden concerning a "population crisis" which was feared as a result of low birth rates at the time.

III. THE PERIOD AFTER 1950

7. The beginnings of a new era. The greatly improved international communications after the end of World War II contributed to making the statistical community in Sweden aware of the recent advancements in sample survey theory, methods, and applications in the United States and India, to mention two of the leading countries. The new developments were studied and discussed, for example, at the conference of the Scandinavian statisticians in Helsinki in 1949. Statisticians were proud to be able to "talk sample survey methods"; to be sure, in some cases this ability was limited to knowledge of certain technical terms, notably "stratification". Mention should also be made of the influence exercised by the United Nations and affiliated agencies such as the Food and Agriculture Organization. In the following we give some examples of sample surveys and related developments of methods and theory. For cases dating to the early 1950's, details are found in Dalenius (1957).
8. The 1950 sample inventory of acreages and livestock. In the 1930's, sample surveys were used to estimate acreages of various crops and animal stocks. These surveys were referred to as "representative counts". They were based on nonprobability selection of farms. The aim, which however was not achieved, was to select 1/10 of the farms in each of several size-groups into which the farms had been divided. In the 1940's, these surveys were carried out on a total basis. A decision was made for the 1950 survey to return to sampling. The design that was suggested and largely implemented for the 1950 survey represented a partial break with the classical tradition of selecting every tenth unit. While the total sample size was fixed by the government authorities to be 1/10 of the total number of farms in the target population, the new design called for stratifying the farms by size groups based on acreage and using minimum variance allocation, which implied a selection of relatively speaking more large farms than small farms. It is interesting to note that the government authorities responsible for assessing the design felt it necessary to consult the U.N. Subcommittee on Statistical Sampling about the appropriateness of the drastic deviation from the "every tenth unit rule". The Subcommittee wholeheartedly endorsed the design. Consequently it was accepted in principle. The design provided considerable opportunity for research. In fact, three contributions to the theory of stratified sampling emerged, namely, (i) how best to divide a population into L strata; (ii) the best choice of the number of strata; and (iii) sample allocation to the strata for estimation of several parameters. The suggested design also called for addressing the problem of "measurement errors" in the acreage, and a special calibration survey was proposed. However, the authorities rejected this proposal.

9. Yield estimation. During World War II, the yield of various crops was estimated using data collected by "eye estimates" of the yield per unit area. By 1950, it was realized that this data collection method could be seriously biased. In the beginning of the 1950's, time was ripe for considering a different approach, namely, crop estimation based on harvesting sample plots, referred to as "objective crop estimation". Accordingly, a pilot study was carried out to test the use of this approach. The outcome of the test was convincing. From then on, the "objective" method has been used. As part of the pilot survey design, a scheme was developed for without replacement selection of $n = 2$ farms from a stratum with probabilities proportional to size, as discussed in Dalenius (1953). The scheme called for dividing each stratum at random into two parts, and selecting one farm from each part.
10. Developments relating to nonsampling errors. In the early 1950's, the problem of non-response received considerable attention in Sweden as in other countries. Surveys with 20-30% nonresponse were not unusual. This generated a vivid and sometimes heated debate in the statistical community about the distortion of the estimates. For a while, the statisticians seemed to have the problem under their control. The public concern about invasion of privacy has lately changed this picture; nonresponse has again become a serious problem. In the last 15 years, several contributions were made in the area of control of nonsampling errors. The problem of "evasive answer bias", to use the term introduced by S. Warner in connection with randomized response, was addressed in Swensson (1976). And Lyberg (1981) successfully tackled the problem of controlling the coding operation in a population census or in a survey with interviews.
11. Respondent burden. In recent years there has been a growing concern about respondent burden and its negative effects on response rates. For example, the target population in many business surveys is the same, rather limited population. The problem can be alleviated by special sample selection techniques. The SAMU system for business surveys at Statistics Sweden permits "negative coordination" of samples, in the sense that samples without overlap can be selected with the technique known as JALES. To each unit in the sampling frame, a uniformly distributed random number is attached. This number stays with the unit, and is used in the selection of samples over time.
12. Modeling in combination with traditional probability sampling principles. Since the 1950's, the methodology for surveys had closely followed the strong probability sampling tradition established by Neyman and by Hansen and his co-workers in the United States. However, sometimes modeling is necessary in surveys when the traditional probability sampling theory is not sufficient. Since the 1970's the use of modeling in surveys has been explored. The book *Foundations of Inference in Survey Sampling* by Cassel, Särndal and Wretman (1977) exposed the new trends. Also, a number of papers by these and other Swedish authors showed how models may assist in inference from surveys. In recent years, methodologists at Statistics Sweden have shown unusual openness to incorporating modeling in the making of survey estimates. An early example where design-based and model-based ideas were combined is the "Öresund survey" for measuring traffic flow between Sweden and Denmark. The design is discussed in Cassel (1978). Some surveys are now designed with the aid of modeling assumptions, as in the work force survey described in Lundström (1987) and in an ongoing project of restructuring of the business survey sector.
13. Safeguarding privacy in surveys. In the last two decades, the general public has become increasingly concerned about invasion of privacy in connection with surveys, including population censuses, carried out by Statistics Sweden. As a result, there has been a trend

towards increasing nonresponse rates in some surveys. Several measures have been taken to deal with the problem: (i) Statistics Sweden has adopted the Ethical Declaration of the International Statistical Institute (1986); a translation of that declaration was distributed to all employees; (ii) In 1987, Statistics Sweden held an international conference which focused on policy issues (as distinguished from "techniques"); the discussions at the conference are summarized in Statistics Sweden (1987); (iii) Statistics Sweden has promoted the development of new safeguards for privacy in its surveys and has taken active steps to apply them. A review is given in Dalenius (1988). Of special interest are papers by Block and Olsson (1976), who describe a measure for the identifying power of quasi-identifiers, and Cassel (1976), who discussed probability-based disclosure.

14. Specific events. The increasing appreciation of sample surveys since around 1950 led to the creation of the Survey Research Center at Statistics Sweden in 1953. A similar interpretation may be given to the establishment of a professorship in "statistics, especially official statistics" at the University of Stockholm in 1965. Also, professorships in survey methodology were recently created at Statistics Sweden.

REFERENCES

- BLOCK, H., and OLSSON, L. (1976). Bakvägsidentifiering. (Backwards identification) *Statistisk Tidskrift*, 1976, 135-144.
- CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.
- CASSEL, C.M. (1978). Probability based disclosure. In Dalenius, T., and Klevmarken, A. (eds.), *Proceedings of a symposium on personal integrity and the need for data in the social sciences*. Swedish Council for Social Science Research, Stockholm, 189-193.
- CASSEL, C.M. (1978). On errors in the predictions with logit models. Technical report, Statistics Sweden.
- DALENIUS, T. (1953). Något om metoder för objektiva skördeberäkningar. (About methods for objective crop estimation.) *Kungliga Lantbruksakademiens Tidskrift*, 92, 99-118.
- DALENIUS, T. (1957). *Sampling in Sweden. Contributions to the Methods and Theory of Sample Survey Practice*. Stockholm: Almqvist and Wiksell.
- DALENIUS, T. (1988). Controlling Invasion of Privacy in Surveys. Statistics Sweden.
- INTERNATIONAL STATISTICAL INSTITUTE (1986). Declaration of Professional Ethics. *International Statistical Review*, 54, 227-242.
- LUNDSTRÖM, S. (1987). Utveckling av estimatorer för skattning av antal förvärvsarbete i olika arbetstidsklasser inom små redovisningsgrupper. R&D Report, U/STM 40, Statistics Sweden.
- LYBERG, L. (1981). Control of the coding operation in statistical investigations. Urval no. 13, Statistics Sweden.
- STATISTICS SWEDEN (1987). Statistics and Privacy: Future Access to Data for Official Statistics – Cooperation or Distrust? Statistics Sweden.
- SWENSSON, B. (1977). Survey measurement of sensitive attributes. Ph.D. Thesis, Department of Statistics, University of Stockholm.

Variance Estimation when a First Phase Area Sample is Reestratified

PHILLIP S. KOTT¹

ABSTRACT

This paper proposes an unbiased variance estimation formula for a two-phase sampling design used in many agricultural surveys. In this design, geographically defined primary sampling units (PSUs) are first selected via stratified simple random sampling; then secondary sampling units within sampled PSUs are reestratified based on their characteristics and subsampled in a second phase of stratified simple random sampling.

KEY WORDS: Two-phase sample; Primary sampling unit; Secondary sampling unit; Unbiased.

1. INTRODUCTION

Suppose we have a sample of geographically defined primary sampling units (PSUs) drawn from a stratified area frame. Each sampled PSU contains a number of secondary sampling units (SSUs) which are reestratified based on their characteristics. Subsamples of the SSUs are then drawn within each new stratum. To avoid confusion, only the original area strata will hereafter be referred to as strata; the new strata based on SSU characteristics will be referred to as *domains*. Stratified simple random sampling (srs) without replacement is performed at both phases of the sampling design.

This article derives an unbiased variance formula for the estimation strategy described above which is used in many agricultural surveys (for example, see Kott and Johnston 1988) but is not restricted to such surveys. The formula is a generalization of a suggestion by Cochran and Huddleston (1969, 1970), who assumed unstratified srs in the first sampling phase. It is also a special case of a variance formula in Särndal and Swensson (1987). The Särndal and Swensson formula (their equation (4.4)) depends on the calculation of a joint inclusion probability for each pair of subsampled SSUs. This proves cumbersome for the particular application under study because there are six distinct situations which need to be considered (depending on whether or not the two SSUs come from the same PSU, stratum, and/or domain). The derivation presented here follows a different line of reasoning entirely.

2. PRELIMINARIES

Suppose we start with an area survey consisting of n_h (out of N_h) PSUs from each of H strata. The SSUs within sampled PSUs are then reestratified into D domains. Within domain d , m_d (out of M_d) SSUs are subsampled. Both phases of the sampling design are stratified srs without replacement.

Let us concentrate on estimating the total for a particular item of interest. To this end, let

¹ Phillip S. Kott, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, DC 20250, USA.

- S^1 = denote the set of all SSUs within a PSU selected in the first phase of sampling whether these SSUs are in the subsample or not,
 S_{hj} = denote the set of subsampled SSUs in PSU j of stratum h ,
 S_h = denote the set of all subsampled SSUs in stratum h ,
 R_d = denote the set of all subsampled SSUs in domain d ,
 x_i = denote the value of interest for SSU i ,
 $e_i = (N_h/n_h)(M_d/m_d)x_i$ (assuming $i \in S_h \cap R_d$) be the “fully expanded” value of interest for SSU i ,

$$e_{dhj} = \sum_{i \in S_{hj} \cap R_d} e_i,$$

$$e_{dh\cdot} = \sum_{i \in S_{h\cdot} \cap R_d} e_i,$$

$$e_{d\cdot\cdot} = \sum_{i \in R_d} e_i,$$

$$e_{\cdot hj} = \sum_{i \in S_{hj}} e_i, \text{ and}$$

$$e_{\cdot h\cdot} = \sum_{i \in S_{h\cdot}} e_i.$$

Note that when S_{hj} is empty, e_{dhj} and $e_{\cdot hj}$ are zero. Likewise when $S_{h\cdot}$ is empty, $e_{dh\cdot}$ and $e_{\cdot h\cdot}$ are zero, and when R_d is empty e_{dhj} , $e_{dh\cdot}$, and $e_{d\cdot\cdot}$ are zero.

An unbiased estimator for X , the sum of x_i values across all SSUs in the population, is

$$\hat{X} = \sum_{d=1}^D \sum_{i \in R_d} e_i. \quad (1)$$

To see this, observe that $\tilde{X} = \sum_{i \in S^1} (N_D/n_D)x_i$ is an unbiased estimator of X with respect to the first phase of sampling, while \hat{X} is an unbiased estimator of \tilde{X} with respect to the second sampling phase. Mathematically, $E_1(\tilde{X}) = X$ and $E_2(\hat{X}) = \tilde{X}$, which implies $E(\hat{X}) = E_1 E_2(\hat{X}) = X$.

3. VARIANCE OF \hat{X}

From any of a number of textbooks on sampling theory (e.g., Cochran 1977, p. 276), we know that the variance of a two-phase estimator like \hat{X} is

$$\text{var}(\hat{X}) = \text{var}_1[E_2(\hat{X})] + E_1[\text{var}_2(\hat{X})], \quad (2)$$

where E_k and var_k denote, respectively, expectation and variance with respect to the k^{th} phase of sampling.

The first term in equation (2) is often called the first phase variance because it equals the variance that would be obtained if every SSU within a sampled PSU were part of the subsample. The second term in (2) is often called the second phase variance. It is easier to estimate than the first phase variance and we will attack it first. The problem with first phase variance estimation is that total value of interest for a PSU in the first phase sample can only be estimated using the subsample. As is well known, putting an estimated PSU total in place of a real total in the usual one-phase variance formula biases the resulting estimator.

3.1 Second Phase Variance Estimation

An unbiased estimator of $\text{var}_2(\hat{X})$ given *any* original sample is automatically an unbiased estimator of $E_1[\text{var}_2(\hat{X})]$. To see this, suppose that v_2 is an unbiased estimator of $\text{var}_2(\hat{X})$ given any sample. Since $E_2[v_2 - \text{var}_2(\hat{X})] = 0$ for *every possible* S^1 , the first phase expectation of $E_2[v_2 - \text{var}_2(\hat{X})]$ must also be zero. Consequently, $E(v_2) = E_1E_2(v_2) = E_1[\text{var}_2(\hat{X})]$.

Now given our particular S^1 ,

$$\hat{\text{var}}_2 = \sum_{d=1}^D (1 - m_d/M_d) [m_d/(m_d - 1)] \left[\left\{ \sum_{i \in R_d} e_i^2 \right\} - e_{d..}^2/m_d \right] \tag{3}$$

is the conventional unbiased estimator for $\text{var}_2(\hat{X})$. Moreover, equation (3) would hold whatever first phase sample obtained. As a result, $\hat{\text{var}}_2$ is also an unbiased estimator for $E_1[\text{var}_2(\hat{X})]$.

3.2 First Phase Variance Estimation

Consider a PSU j within stratum h . The value $e_{.hj}$ is an unbiased estimator of (N_h/n_h) times the total value among all SSUs in PSU j whether in the current subsample or not. Consequently, $E_2(e_{.hj})$ is exactly equal to (N_h/n_h) times the total value among all SSUs in PSU j . With this in mind, the following would be an unbiased estimator of the first phase variance of \hat{X} :

$$\begin{aligned} \hat{\text{var}}_1[E_2(\hat{X})] = \\ \sum_{h=1}^H (1 - n_h/N_h) [n_h/(n_h - 1)] \left[\sum_{j=1}^{n_h} \{E_2(e_{.hj})\}^2 - \{E_2(e_{.h.})\}^2/n_h \right]. \end{aligned} \tag{4}$$

Taken as is, equation (4) is of little use since it supposes we know what the $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ are. Nevertheless, it does suggest that $\text{var}_1[E_2(\hat{X})]$ would be estimated in an unbiased manner if one could find unbiased estimators for the $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ to plug into (4).

Observe first that $e_{.hj}^2$ and $e_{.h.}^2$ are *not* unbiased estimators of $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$. In fact,

$$E_2(e_{.hj}^2) = \{E_2(e_{.hj})\}^2 + \text{var}_2(e_{.hj}), \tag{5}$$

while

$$E_2(e_{.h.}^2) = \{E_2(e_{.h.})\}^2 + \text{var}_2(e_{.h.}).$$

These equations hint towards alternative estimators for $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$. If v_{2hj} and $v_{2h.}$, say, were unbiased estimators of $\text{var}_2(e_{.hj})$ and $\text{var}_2(e_{.h.})$, respectively, then $e_{.hj}^2 - v_{2hj}$ would be an unbiased estimator of $\{E_2(e_{.hj})\}^2$, while $e_{.h.}^2 - v_{2h.}$ would be an unbiased estimator of $\{E_2(e_{.h.})\}^2$.

From Cochran (1977, p. 143, eq. (5A.68)), one can see that

$$\text{vâr}_{2hj} = \sum_{d=1}^D (1 - m_d/M_d) [m_d/(m_d - 1)] \left[\left\{ \sum_{i \in S_{hj} \cap R_d} e_i^2 \right\} - e_{dhj}^2/m_d \right]$$

and

$$\text{vâr}_{2h.} = \sum_{h=1}^H (1 - m_d/M_d) [m_d/(m_d - 1)] \left[\left\{ \sum_{i \in S_{h.} \cap R_d} e_i^2 \right\} - e_{dh.}^2/m_d \right] \quad (6)$$

are, respectively, unbiased estimators of $\text{var}_2(e_{.hj})$ and $\text{var}_2(e_{.h.})$.

3.3 Putting It All Together

Observe that combining equations (3) and (6) can yield (after some manipulation) this estimator for the second phase variance of \hat{X} :

$$\begin{aligned} \text{vâr}_2 = & \sum_{h=1}^H [n_h/(n_h - 1)] \sum_{j=1}^{n_h} \text{vâr}_{2hj} - \text{vâr}_{2h.}/(n_h - 1) + \\ & \sum_{d=1}^D \left\{ (1 - m_d/M_d) [1/(m_d - 1)] \cdot \right. \\ & \left. \left(\sum_{h=1}^H [n_h/(n_h - 1)] \left[\left\{ \sum_{j=1}^{n_h} e_{dhj}^2 \right\} - e_{dh.}^2/n_h \right] - e_{d.}^2 \right) \right\}. \end{aligned} \quad (7)$$

By plugging $e_{.hj}^2 - \text{vâr}_{2hj}$ and $e_{.h.}^2 - \text{vâr}_{2h.}$ respectively into $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ of equation (4), we have the following estimator for the first phase variance of \hat{X} :

$$\begin{aligned} \text{vâr}_1[E_2(\hat{X})] = & \sum_{h=1}^H (1 - n_h/N_h) [n_h/(n_h - 1)] \cdot \\ & \left[\left\{ \sum_{j=1}^{n_h} e_{.hj}^2 - \text{vâr}_{2hj} \right\} - \{ (e_{.h.}^2 - \text{vâr}_{2h.}) \}/n_h \right]. \end{aligned}$$

This can then be added to (7) to yield the following estimator for the variance of \hat{X} in (1):

$$\text{vâr} = A + B + C, \quad (8)$$

where

$$A = \sum_{h=1}^H [n_h / (n_h - 1)] \left[\left\{ \sum_{j=1}^{n_h} e_{.hj}^2 \right\} - e_{.h.}^2 / n_h \right],$$

$$B = \sum_{d=1}^D \left\{ (1 - m_d / M_d) [1 / (m_d - 1)] \cdot \left(\sum_{h=1}^H [n_h / (n_h - 1)] \left[\left\{ \sum_{j=1}^{n_h} e_{dhj}^2 \right\} - e_{dh.}^2 / n_h \right] - e_{d.}^2 \right) \right\},$$

$$C = - \sum_{h=1}^H f_h n_h / (n_h - 1) \left[\sum_{j=1}^{n_h} \{ e_{.hj}^2 - \hat{v}ar_{2hj} \} - \{ e_{.h.}^2 - \hat{v}ar_{2h} \} / n_h \right],$$

$f_h = n_h / N_h$ is the first phase sampling fraction in stratum h , and $\hat{v}ar_{2hj}$ and $\hat{v}ar_{2h}$ are defined by equation (6).

Observe that if all the first phase sampling fractions are very small, then the contribution of C to (8) can be ignored. In any event dropping C would at worst give $\hat{v}ar$ an upward bias, since $E(C) \leq 0$.

Observe further that $\hat{v}ar$ would collapse to A if – in addition to C being ignorably small – the sampling design had been conventional two-stage sampling; that is, if each domain had been contained within one of the originally sampled PSU's so that $y_{d..} = y_{dhj} = y_{dh.}$ and $B = 0$. This should not be surprising, since A is the standard variance estimator in two stage sampling when the first stage is srs with replacement (Cochran 1977, p. 307). Ignorable first stage sampling fractions blur the distinction between srs with and without replacement.

The right hand side of (8) can, in principle, be negative. This is because B is often negative (since $y_{d..} \geq y_{dh.} \geq y_{dhj}$), while A can theoretically be as small as zero. Kott and Johnston (1988) applied a formula similar to (6) to data from a US Department of Agriculture survey. In the 41 cases they examined the absolute value of B was always less than 7% of A .

One final note. Since $B \leq 0$ and $E(C) \leq 0$, using A alone provides a conservative, unambiguously nonnegative, estimate for $\text{var}(\hat{X})$.

REFERENCES

- COCHRAN, R., and HUDDLESTON, H. (1969). Unbiased estimates for stratified subsample designs. U.S. Department of Agriculture, Statistical Reporting Service.
- COCHRAN, R., and HUDDLESTON, H. (1970). Unbiased estimates for stratified subsample design. *Proceedings of the Section on Social Statistics, American Statistical Association*, 265-267.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3rd. ed.). New York: Wiley.
- KOTT, P.S., and JOHNSTON, R. (1988). Estimating the non-overlap variance component for multiple frame agricultural surveys. RAD Staff Report No. SRB-NERS-8805, U.S. Department of Agriculture, National Agricultural Statistics Service.
- SÄRNDAL, C.E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Estimation Using Double Sampling and Dual Stratification

DONALD B. WHITE¹

ABSTRACT

The problem considered is that of estimation of the total of a finite population which is stratified at two levels: a deeper level which has low intrastratum variability but is not known until the first phase of sampling, and a known pre-stratification which is relatively effective, unit by unit, in predicting the deeper post-stratification. As an important example, the post-stratification may define two groups corresponding to responders and non-responders in the situation of two-phase sampling for non-response. The estimators of Vardeman and Meeden (1984) are employed in a variety of situations where different types of prior information are assumed. In a general case, the standard error relative to that of the usual methods is studied via simulation. In the situation where no prior information is available and where proportional sampling is employed, the estimator is unbiased and its variance is approximated. Here, the variance is always lower than that of the usual double sampling for stratification. Also, without prior information, but with non-proportional sampling, using a slight modification of the second phase sampling plan, an unbiased estimator is found along with its variance, an unbiased estimator of its variance, and an optimal allocation scheme for the two phases of sampling. Finally, applications of these methods are discussed.

KEY WORDS: Two-phase sampling; Prior information; Variance estimation; Optimal allocation; Non-response.

1. INTRODUCTION

Various stratified sampling designs employ various types of prior information. For example, the usual stratification model assumes full prior knowledge of individual stratum memberships. Post-stratification is useful when there is global information on stratum sizes but no information on individuals. Double sampling for stratification, on the other hand, assumes no prior information on strata. Further, some knowledge of the population values is necessary, for example, for the allocation of sampling resources among strata (see, for example, Cochran 1977, pp. 96-99 and 331-332).

The rigid assumptions inherent to these sampling designs and population models often are not satisfied due to the discrepancy between the population under study and the (possibly dated) prior information. Seeking to appropriately handle this discrepancy, Vardeman and Meeden (1984) have introduced a pair of estimators which combine information on stratum memberships, stratum sizes, and stratum averages with analogous information gained from the current sample. Their two estimators apply to two essentially different situations. The first is where the prior information is global only, *i.e.*, only on stratum sizes and averages. The second estimator applies where there is also partial information on individual stratum memberships. Here, the population is stratified according to various factors, some of which are known and some of which, though not known, may be inexpensive to determine on a first phase of sampling.

¹ Donald B. White, Department of Statistics, State University of New York at Buffalo, 249 Farber Hall, Buffalo, New York 14214.

As an example, consider the use of sampling to determine the spread of an infectious disease. If detection of infection is expensive, then stratification, according to risk categories, is desirable to reduce the second phase sample size. Factors determining risk categories may include gender, age, place of residence, ethnicity, health habits, and contact with potential carriers. As some of these factors are not known prior to sampling, the model of Vardeman and Meeden can be employed since the true risk categories can be predicted by the known factors.

Another example is two-phase sampling for non-response. Extending the method of Hansen and Hurwitz (1946), we have a population which is divided into two post-strata, *i.e.*, responders and non-responders. The methods discussed here apply when there is some prior information which classifies units into pre-strata which are then used to predict whether or not the unit will be in the group of responders.

The notion of employment of prior information in two-phase designs is not without precedent in the sampling literature. As an example, Han (1973) has used prior information on an auxiliary regression variable (to be measured in a first phase sample) to construct a simple hypothesis (say H_0) regarding the mean of that variable. The first phase sample measurements are then used to test H_0 . If H_0 is accepted, the value specified by H_0 is used in the estimator; if it is rejected, the sample average is used.

A discussion of the use of the first estimator of Vardeman and Meeden (global information only) can be found in White (1987). There, optimal choices of the weighting constants for prior information relative to the information contained in the current sample were determined. Here, the situation considered is where prior information is also available on individual stratum memberships. After introducing the necessary notation in Section 2, we explore a simulated example in Section 3. In Section 4, in two different sampling situations, unbiased estimators are analyzed in terms of variance, unbiased estimation of the variance, and optimal allocation of sampling resources. In Section 5, applications of these techniques are discussed.

2. THE POPULATION MODEL AND SAMPLING SCHEME

We now present the population model and the proposed sampling design. We begin with a finite population P of units labelled $1, 2, \dots, N$ with associated unknown values y_1, y_2, \dots, y_N . Denote the population total by $\tau = \sum_{i=1}^N y_i$. For $1 \leq i \leq N$, unit i also possesses an unknown post-stratum membership j_i , $1 \leq j_i \leq J$, and a known pre-stratum membership k_i , $1 \leq k_i \leq K$.

A variety of population quantities require a specialized notation. Such quantities include sizes of groups, group averages and group variances. Subscripts will identify the group involved: no subscript implies reference to the entire population, " k ." refers to pre-stratum k , $1 \leq k \leq K$, " j " refers to post-stratum j , $1 \leq j \leq J$, and the subscript " kj " refers to the intersection of pre-stratum k with post-stratum j . The base symbols N , \bar{Y} and S^2 refer to number of elements, y -average, and finite population variance, respectively. Also, we let P , $P_{k\cdot}$, $P_{\cdot j}$ and P_{kj} denote the subsets of P corresponding to the four categories given above. For example, we have

$$S_{k\cdot}^2 = \frac{1}{N_{k\cdot} - 1} \sum_{i \in P_{k\cdot}} (y_i - \bar{Y}_{k\cdot})^2.$$

Also, we can write

$$\tau = \sum_j N_{.j} \bar{Y}_{.j}. \quad (1)$$

We finally let $W_{kj} = N_{kj}/N_{k.}$, i.e., W_{kj} is the proportion of units in pre-stratum k which fall into true stratum j .

We now discuss the sampling technique. In the first phase of sampling, a stratified simple random sample without replacement s' is selected, with n'_k units (first phase sampling fraction denoted by $f'_{k.} = n'_{k.}/N_{k.}$) selected from pre-stratum k . Samples from different pre-strata are independent. For these $n' = \sum_k n'_{k.}$ units, post-strata, j_i , are observed. Following the notational pattern given above, we let n'_{kj} denote the number of units in s' sampled from pre-stratum k which happen to fall in post-stratum j . Also, $n'_{.j} = \sum_k n'_{kj}$ is the total number of units in s' which fall in post-stratum j . This set of units is denoted by $s'_{.j}$. These quantities are observed, while quantities involving y -values, such as \bar{y}' and s'^2 (with all four types of subscripts), remain unobserved. Here, and in the following, the average of any empty collection is taken as zero, and, if the size of a group is one or zero, we take its variance s^2 to be zero. We note that for $1 \leq k \leq K$, the random vectors $(n'_{k1}, \dots, n'_{kj})$ are independent with each possessing a multivariate hypergeometric distribution.

For the second phase of sampling, we partition s' into $\cup_{j=1}^J s'_{.j}$, i.e., by post-stratification. For each j , let $v_j(\cdot)$ denote a known function on and into the non-negative integers with $v_j(0) = 0$ and $1 \leq v_j(x) \leq x$ if $x \geq 1$. The second phase sample s is also stratified, but now is a subsample of s' and stratified according to the post-stratification. The sample from $s'_{.j}$ is denoted $s_{.j}$ and is of size $n_{.j} \equiv v_j(n'_{.j})$. Here, y -values are observed, yielding quantities such as $\bar{y}_{.j}$ and $s_{.j}^2$, the y -average and finite population variance of the units in the phase two sample and stratum j .

The estimates of τ given by Vardeman and Meeden include the option of inclusion of prior guesses for the relative stratum sizes within each pre-stratum and for the stratum averages. Thus, we have prior guesses for the values W_{kj} and $\bar{Y}_{.j}$ which are given by Π_{kj} and $\mu_{.j}$, respectively. In the estimator introduced below, these guesses are given weighting constants which reflect the confidence in the guess relative to the confidence in the corresponding information yielded by the current sample. For each k , the confidence value allotted to the collection $(\Pi_{k1}, \dots, \Pi_{kJ})$ is denoted $\tilde{M}_k \in [0, \infty]$ and for each j , the confidence value given to $\mu_{.j}$ is denoted $M_{.j} \in [0, \infty]$. In the current sample, the collection (W_{k1}, \dots, W_{kJ}) is estimated by $(n'_{k1}/n'_{k.}, \dots, n'_{kJ}/n'_{k.})$ and is based on a simple random sample of size $n'_{k.}$. Thus, the confidence in Π_{kj} , say, as opposed to $n'_{kj}/n'_{k.}$, is reflected by the size of \tilde{M}_k versus that of $n'_{k.}$. Similarly, in the current sample, $\bar{Y}_{.j}$ is estimated by $\bar{y}_{.j}$ and is based on a sample of size $n_{.j}$; thus, the relative confidence in the prior guess and the current estimate is reflected by the relative sizes of $M_{.j}$ and $n_{.j}$. Any confidence weight for prior information equal to zero corresponds to no use of the prior information, and, as in the use of stratum sizes in the usual post stratification model, a value of infinity implies no use of the corresponding information in the current sample.

Using the prior guesses, current estimates and confidence weights, we estimate W_{kj} and $\bar{Y}_{.j}$ by $\hat{\Pi}_{kj} = (\tilde{M}_k \Pi_{kj} + n'_{kj})/(\tilde{M}_k + n'_{k.})$ and $\hat{\mu}_{.j} = (M_{.j} \mu_{.j} + n_{.j} \bar{y}_{.j})/(M_{.j} + n_{.j})$, respectively. Finally, an estimate $\hat{\tau}$ of the population total τ is constructed by replacing in the formula (1) for τ any unobserved quantity by its estimate given above. Thus, we employ

$$\hat{\tau} = \sum_{j=1}^J \left\{ n_{.j} \bar{y}_{.j} + (n'_{.j} - n_{.j}) \hat{\mu}_{.j} + \sum_{k=1}^K (N_{k.} - n'_{k.}) \hat{\Pi}_{kj} \hat{\mu}_{.j} \right\}. \quad (2)$$

Computation of the bias and variance of $\hat{\tau}$ in the general case is left open by Vardeman and Meeden. The case $K = 1$ and $M_{.j} = 0, 1 \leq j \leq J$, has been studied in White (1987). Before proceeding to a result in a more complex situation, we first explore the results of a simulation on a hypothetical population.

3. A MONTE CARLO STUDY

Here we present a specific population and sampling scheme which is modelled after the introductory example regarding estimation of the spread of an infectious disease. For a population of 10,000 individuals who are susceptible, the disease is assumed to be more prevalent among the 5,000 who live in the western section of the area considered. Since this is a known characteristic, the population is partitioned according to the east-west boundary into $K = 2$ pre-strata. Next, we assume that certain easily obtained additional information enables the sampler to categorize the individual as low, medium, or high risk for becoming infected. See Table 1 for the details of the construction of the population.

For estimation of the total number infected ($\tau = 2302$), we assume no prior knowledge of the stratum proportions $\bar{Y}_{.1}, \bar{Y}_{.2},$ and $\bar{Y}_{.3}$ and thus take $M_{.1} = M_{.2} = M_{.3} = 0$. There remain four major ingredients to the estimation process: 1) the prior guesses $\{\Pi_{kj}; k = 1, 2, j = 1, 2, 3\}$ for the distribution of individuals from pre-strata to post-strata, 2) the weighting constants \tilde{M}_1 and \tilde{M}_2 given to these prior guesses, 3) the first phase sample design and outcome, and 4) the second phase sample design and outcome. These are detailed in the following.

First, in White (1987) it was found for the $K = 1$ case that an effective choice of weighting constants was to select M equal to the sample size on which the previous information was based. Following that notion, we allowed, for each simulation, the collection $\{\Pi_{kj}\}$ to select itself through a preliminary sample of size m (either 500 or 2500) from each pre-stratum. That is, Π_{kj} is taken to be the proportion of the m individuals from pre-stratum k falling in post-stratum j .

Second, for each run, the weighting constants were taken as $\tilde{M}_1 = \tilde{M}_2 = M$ for all $M \in \{0, 100, 200, 300, \dots, 10,000, \infty\}$. Recall that $M = \infty$ corresponds to the situation of the usual post-stratification where no use is made of the current sample to estimate group sizes.

Third, the first phase sample is stratified according to pre-strata with sampling fractions f'_k taken to be $f'_1 = f'_2 = f, f \in \{.10, .20, .30, .40, .50\}$. Recall that in this phase of sampling, only post-stratification is observed. This information is, presumably, inexpensive to obtain.

Table 1
Number Infected/Group Size for the Pre-strata and Post-strata Combinations

Location of Residence	Risk Group j	Low 1	Medium 2	High 3	Total
East ($k = 1$)		40/4000	80/800	100/200	220/5000
West ($k = 2$)		2/200	80/800	2000/4000	2082/5000
Total		42/4200	160/1600	2100/4200	2302/10000

On the other hand, sampling a unit in phase 2, where the presence of infection is determined, is assumed to be rather expensive. The individuals selected are a subsample of the phase one sample, stratified according to post-strata. The sampling fractions in various strata are again taken as equal ($v_j(n'_{.j}) = [c_j n'_{.j}]$ for $n'_{.j}$ large enough, and $c_1 = c_2 = c_3 = c$) and so that different simulations can be compared, c is selected so that the fraction of the entire population which appears in the phase 2 sample remains constant at .10.

Now, the following process is repeated $R = 50,000$ times: obtain a preliminary sample of size m from which prior guesses Π_{kj} for W_{kj} are constructed. Next, a sample, stratified according to pre-strata with sampling fractions f , is obtained. Only post-stratification is observed. Then, a subsample, stratified according to post-strata with sampling fractions c , is obtained and units in this sample are classified as infected or not infected. Finally, on each run, $\hat{\tau}$ is obtained for each value of M considered. The standard error of $\hat{\tau}$ is estimated using the R simulated values of $\hat{\tau}$. Recall, however, that in a real-life application, the standard error of an estimate will depend on the particular values of Π_{kj} used; here, these values are different on each run and thus the estimated standard error should be viewed as a long run average for a mixture of distributions of $\hat{\tau}$, mixed according to the distribution of the Π_{kj} based on the preliminary sample.

The simulations were performed on an IBM3031 computer. For this example, where $y_i \in \{0,1\}$ for all i , all random quantities are functions of independent hypergeometric or multivariate hypergeometric variables. Using the fact that the conditional distribution of a univariate marginal of a multivariate hypergeometric distribution given any subcollection of the other coordinates is itself hypergeometric, all random quantities were simulated using the IMSL 92DP hypergeometric simulation subroutine GGHPR. For the first combination of m and f (500 and .10), the simulation process was repeated five times to check internal consistency.

Tables 2 and 3 summarize pertinent characteristics of the variation of the simulated $SE(\hat{\tau})$ as a function of M for the five repeated simulations (Table 2), and the simulations for various values of f and m (Table 3). Table 2 gives only highlights which demonstrate internal consistency and confirm that the number of repetitions is chosen large enough. Note that M_0 denotes the value of M for which $SE(\hat{\tau})$ is minimized. In Table 3, also given is a comparison with the better of the possible usual techniques (regular two-phase or stratified according to pre-strata) relative to the ideal where the true strata are regarded as known. The standard error of an estimator based on stratified sampling using pre-strata only is 113.27, and for stratified according to true strata, it is 105.47. Thus, letting the estimator in regular two-phase sampling be denoted by $\hat{\tau}_2$ and realizing that $SE(\hat{\tau}_2)$ depends upon f and c , the values appearing in the columns headed Percent Relative Reduction in $SE(\hat{\tau})$ are $100 [\min(SE(\hat{\tau}_2), 113.27)] - SE(\hat{\tau}) / [\min(SE(\hat{\tau}_2), 113.27) - 105.47]$.

Table 2
Key Features of the Repeated Runs with $m = 500, f = .10$ and $c = 1.0$

Run #	M_0	SE($\hat{\tau}$)			
		$M = 0$	$M = m$	$M = M_0$	$M = \infty$
1	600	113.55	109.67	109.62	112.00
2	700	113.42	109.50	109.45	111.80
3	700	113.92	109.86	109.78	112.00
4	600	113.61	109.71	109.66	112.07
5	600	113.56	109.74	109.70	112.17

Table 3
Key Features of SE($\hat{\tau}$) as a Function of M

m	f'	c	SE($\hat{\tau}_2$)	M_0	SE($\hat{\tau}$)				Percent Relative Reduction in SE($\hat{\tau}$)		
					$M = M_0$	$M = 0$	$M = m$	$M = \infty$	$M = 0$	$M = m$	$M = \infty$
500	.10	1.00	126.29	600	109.62	113.55	109.67	112.00	-3.6	46.2	16.3
500	.20	.50	115.19	600	107.95	109.02	107.97	110.72	54.5	67.9	32.7
500	.30	.33	111.80	600	107.87	108.25	107.87	110.38	56.1	62.1	22.4
500	.40	.25	109.22	750	106.51	106.76	106.52	108.29	65.6	72.0	24.8
500	.50	.20	107.98	700	106.17	106.28	106.18	107.55	67.7	71.7	17.1
2500	.10	1.00	126.29	*	\leq 106.20	113.33	106.42	106.20	-0.8	87.8	90.6
2500	.20	.50	115.19	*	\leq 105.76	108.67	106.02	105.76	59.0	92.9	96.3
2500	.30	.33	111.80	*	\leq 106.63	108.18	106.87	106.63	57.2	77.9	81.7
2500	.40	.25	109.22	*	\leq 105.77	106.59	105.94	105.77	70.1	87.5	92.0
2500	.50	.20	107.98	*	\leq 105.81	106.34	105.96	105.81	65.3	80.5	86.5

* -- $> 10,000$

A variety of important results can be discerned from Table 3. First is that for $m = 500$, M_0 is very close to, although always slightly larger than, m . This is the result predicted by the $K = 1$ situation from White (1987). For $m = 2500$, though in every case $M_0 > 10,000$, one discovers that SE($\hat{\tau}$) at $M = m$ is very close to the minimum at $M = M_0$.

Second is that at $M = m$, the percent relative reduction in SE($\hat{\tau}$) ranges from a minimum of 46% to over 90%. Also, at $M = 0$, corresponding to the situation of dual stratification with no prior information on any population characteristic, the percent relative reduction in SE($\hat{\tau}$) is always over 50% except in the case of the smallest first phase sampling fraction, $f = .10$. In that case, when prior information is not available and the first phase sample size is small, one is better off to use the pre-strata and ignore the true stratification. On the other hand, if one does have a set of prior guesses available for the collection of W_{kj} , but is uncertain of what weights to attach to these values, one could use the usual post-stratification notion of using weight $M = \infty$. If the prior information is good, as in our case $m = 2500$, then the percent relative reduction in SE($\hat{\tau}$) is always over 80%. Even if the prior information is only moderately accurate, as in the case $m = 500$, the reduction in standard error is between 16% and 33%.

In summary, if one is able to identify a weighting constant applicable to prior information on the distributions of units among strata, then a substantial reduction in standard error can be obtained using these methods. Even if one cannot identify such a constant or does not have applicable prior information, one can still decrease standard error using dual stratification by taking $M = 0$ if the prior information on W_{kj} is either poor or non-existent, or $M = \infty$ with accurate prior information. In particular, it thus turns out that the case $M = 0$ is important. This case is examined in detail in the next section.

4. BIAS, STANDARD ERROR, AND OPTIMAL ALLOCATION
WITH NO PRIOR INFORMATION

When no prior information is available, we set $M_j = 0$ and $\tilde{M}_k = 0$ for each $1 \leq j \leq J$ and $1 \leq k \leq K$. In this section, we at first also assume that sampling in both phases is proportional to the size of the group from which the sample is drawn, that is, for each

$k, n'_{k.} = fN_{k.}$ (i.e., $f'_{k.} = f$, all k) and for each $j, n_{.j} = cn'_{.j}$ (i.e., $v_j(x) = cx$, all j). This, of course, immediately introduces an approximation (referred to in what follows as approximation A1), since the resulting sample sizes are not necessarily integers. However, in reasonably large populations, and for reasonably large sampling fractions f and c , this approximation has little impact on the derivations that follow.

In this situation, $\hat{\mu}_{.j}$ reduces to $\bar{y}_{.j}$ and $\hat{\Pi}_{kj}$ reduces to $n'_{kj}/n'_{k.}$ and, thus, we have $\hat{\tau} = 1/f \sum_{j=1}^J n'_{.j} \bar{y}_{.j}$. The derivations of the expectation and variance of $\hat{\tau}$ are summarized in the appendix. The key features are two conditioning arguments: first, we condition on s' since the second phase sample is a function of s' and, second, because of the multivariate hypergeometric nature of the phase one sample, we condition on the values n'_{kj} , the sizes of the various pre-stratum and post-stratum combinations in the first phase sample.

In the appendix, we show first that $\hat{\tau}$ in this case is unbiased (aside from approximation A1) and that an approximation of its variance is given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_k N_{k.} S_{k.}^2 + \frac{1-c}{fc} \sum_j N_{.j} S_{.j}^2. \tag{3}$$

As discussed in the appendix in more detail, formula (3) 1) gives answers close to the simulated values, 2) is based on approximations whose error is small for large populations and reasonably large samples, and 3) reduces to the exact formula in all three of the standard situations. In addition, it is easy to show that the variance given by (3) is always smaller than that of the situation of regular two phase sampling.

Now as in any stratification model, there is a question of optimal design. The problem addressed here is that of minimum variance given a fixed cost. To this end, we let $T_1 = \sum_k N_{k.} S_{k.}^2$ and $T_2 = \sum_j N_{.j} S_{.j}^2$. We assume, for the design question at hand, that these are known. In reality, of course, only guesses are available. Next, we let D denote the total budget, d_0 , the start-up cost, d_1 , the cost per unit in the phase one sample, and d_2 , the cost per unit in the phase two sample. Letting D_a denote the number of dollars available for sampling per population unit, we have

$$D_a = \frac{D - d_0}{N} = f(d_1 + cd_2). \tag{4}$$

With f and c subject to constraint (4), we seek to minimize (3), $\text{var}(\hat{\tau})$, now given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} T_1 + \frac{1-c}{fc} T_2. \tag{5}$$

The solution is easily found to be given by

$$c = \left[\frac{d_1 T_2}{d_2 (T_1 - T_2)} \right]^{1/2} \tag{6}$$

with f found using (4). If $T_1 \leq T_2$, we automatically take $c = 1$ since then the pre-stratification is more effective than the post-stratification.

In the case of non-proportional sampling, the estimator given is biased and calculations of the bias and standard error in this more general situation are prohibitive. However, a slight modification of the second phase sampling design along with the associated change in the estimator $\hat{\tau}$ yields an estimator which is unbiased. Following a description of the required modification, we compute the variance and an unbiased estimator of the variance and we find an optimal method of allocating sampling resources to the various pre- and post-strata.

The modification to the sampling plan is to leave the second phase sample within pre-strata rather than pooling within post-strata across pre-strata. Thus, given n'_{kj} units appearing in $s' \cap P_{kj}$, we have a function $v_{kj}(\cdot)$ (like $v_{.j}(\cdot)$ in Section 2) which defines a sample size $n_{kj} = v_{kj}(n'_{kj}) = c_{kj}n'_{kj}$ to be taken by simple random sampling from $s' \cap P_{kj}$. Based upon this sample, we obtain the quantities \bar{y}_{kj} and s_{kj}^2 which were defined in Section 2. The estimator is now $\hat{\tau} = \sum_k 1/f'_{k.} \sum_j n'_{kj} \bar{y}_{kj}$.

Now, since samples (and thus estimators) are independent between pre-strata, $\hat{\tau}$ is the sum of independent estimators of the K pre-stratum totals, where each estimator is based on a regular double sampling scheme. Thus, the results of Rao (1973) apply to each pre-stratum and we first observe that $\hat{\tau}$ is unbiased because its summands are unbiased estimators of their respective pre-stratum totals. Second, using Rao's results, we have

$$\text{var}(\hat{\tau}) = \sum_k \frac{1}{f_{k.}} \left[(N_{k.} - n'_{k.}) S_{kj}^2 + \sum_j N_{kj} S_{kj}^2 (1/c_{kj} - 1) \right]. \quad (7)$$

Also, an unbiased estimator of $\text{var}(\hat{\tau})$ is given by

$$\begin{aligned} \widehat{\text{var}}(\hat{\tau}) = & \sum_k N_{k.} \left[(N_{k.} - 1) \sum_j \left(\frac{n'_{kj} - 1}{n'_{k.} - 1} - \frac{n'_{kj} - 1}{n'_{k.} - 1} \right) \frac{n'_{kj} s_{kj}^2}{n'_{k.} n_{kj}} \right. \\ & \left. + \frac{N_{k.} - n'_{k.}}{N_{k.} (n'_{k.} - 1)} \sum_j \frac{n'_{kj}}{n'_{k.}} \left(\bar{y}_{kj} - \sum_{j'} \frac{n'_{kj'}}{n'_{k.}} \bar{y}_{kj'} \right)^2 \right]. \quad (8) \end{aligned}$$

We note at this point that in the case of proportional sampling considered earlier in this section, we have proposed two different estimators for τ , one based on a pooled second phase sample, the other unpooled. In both cases, the estimator was found to be unbiased, and, also, reduction of formula (7) to the case where $f'_{k.} = f$ for all k and where $c_{kj} = c$ for all k and all j yields formula (3), *i.e.*, the approximate variance for the pooled second phase sampling estimator.

Finally, again following the results in Rao, we derive an optimal allocation of sampling resources. Say that D dollars are available for the two phases of sampling, where sampling a unit in phase 1 from $P_{k.}$ costs $d'_{k.}$ dollars and sampling a unit in phase 2 from $P_{.j}$ costs $d_{.j}$ dollars. Given these costs, we wish to find the values of $f'_{k.}$ and c_{kj} which minimize the variance of $\hat{\tau}$. Using the Cauchy inequality for the phase 2 sample in each pre-stratum, we observe that no matter what the value of $f'_{k.}$, the sampling fraction from post-stratum j is given by

$$c_{kj} = S_{kj} \left(\frac{d'_{k.}}{d_{.j} (S_{k.}^2 - \sum_j w_{kj} S_{kj}^2)} \right)^{1/2}. \quad (9)$$

Now, the effective expected cost (over both phases of sampling) for each unit sampled in phase 1 and in pre-stratum k is given by

$$d_k^{(e)} = d'_k + \sum_j w_{kj} c_{kj} d_j. \quad (10)$$

When viewed in this way, for cost considerations, the first phase of sampling can be seen as a regular stratified sample with (effective) cost of a unit sampled in P_k given by (10). Thus, Cochran (1977,p.97) provides the required formulation of the first phase allocation:

$$\frac{n'_k}{n'} = \frac{N_k \cdot S_k / \sqrt{d_k^{(e)}}}{\sum_{k'} N_{k'} \cdot S_{k'} / \sqrt{d_{k'}^{(e)}}} \quad (11)$$

where

$$n' = \sum_k n'_k = D \sum_k \frac{N_k \cdot S_k / \sqrt{d_k^{(e)}}}{\sum_{k'} N_{k'} \cdot S_{k'} / \sqrt{d_{k'}^{(e)}}} \quad (12)$$

Following the modifications suggested by Rao, one can handle the situation where one or more of the c_{kj} turn out to be greater than one. One can also modify the results in the usual way to minimize sampling cost in the case of pre-determined variance.

5. APPLICATIONS

One can employ the method of dual stratification presented here at two levels. At one level, double sampling with pre-strata can be employed with no use of prior information on stratum sizes or stratum averages. At a more complex level, if one has in hand prior information on the number of units in each stratum coming from each pre-stratum, and if the sampler has a level of confidence for this information, then a further reduction in standard error can be obtained by employing this prior information.

This two phase sampling and estimation technique could be used in the proposed nationwide survey to determine the extent of spread of the HTLV-III (Acquired Immune Deficiency Syndrome) virus. The extended incubation period, estimated to be on the average 4.5 years (Lui *et al.* 1986), makes the survey approach imperative, yet there are psychosocial and financial factors which make such a survey extremely difficult to carry out. Thus, methods which assist in reducing sample size while maintaining accuracy must be pursued.

Allen (1984) provides data which suggests a partition of the American population according to a variety of factors which can be used to define risk categories. Known factors, which could be used to define pre-strata, include age, gender, presence of certain diseases, nationality, immigration status, and geographical location. Unknown factors, which could be determined via interview, include sexual preference and drug use. Data on the prevalence of HTLV-III within various subgroups can be both 1) incorporated into the overall estimate of prevalence and 2) used to determine sampling allocations. Such data is available, for example, for blood donors (Kuritsky *et al.* 1986), military results (Redfield and Burke 1987), intravenous drug

abusers in Queens, New York (Robert-Guroff *et al.* 1986) and male homosexuals in Greenwich Village (Casareale *et al.* 1984/5). Though this prior information can be used to reduce cost and increase accuracy, confidentiality and sensitivity/specificity of the HTLV-III test remain as significant obstacles which must be addressed carefully before such a study will provide meaningful results.

ACKNOWLEDGEMENT

The author would like to express his appreciation to the reviewer for helpful comments in the area of non-proportional sampling.

APPENDIX

Derivation of Expectation and Variance With No Prior Information and Proportional Sampling

Using the notation given in Section 2, we proceed first with the derivation of $E(\hat{\tau})$. The conditional expectation given s' is $E(\hat{\tau} | s') = 1/f \sum_j n'_{.j} \bar{y}'_{.j}$. Then, writing $n'_{.j} \bar{y}'_{.j}$ as $\sum_k n'_{kj} \bar{y}'_{kj}$, we find $E(\hat{\tau}) = E(E(\hat{\tau} | s')) = 1/f \sum_j \sum_k E(n'_{kj} E(\bar{y}'_{kj} | n'_{kj})) = 1/f \sum_j \sum_k E(n'_{kj}) \bar{Y}_{kj} = \tau$ since n'_{kj} is hypergeometric with sampling fraction f and N_{kj} units in pre-stratum k and post-stratum j . Thus, $\hat{\tau}$ is, in this case, unbiased (ignoring approximation A1).

Computation of the variance is along the same lines, yet much more technically detailed. Only certain elements of the computation will be presented and particular emphasis will be placed on the points in the derivation where approximations are made. First, some computation using the two phases of conditioning discussed above, yields

$$\text{var}(E(\hat{\tau} | s')) = \frac{1-f}{f} \sum_k N_k S_k^2. \quad (13)$$

We next obtain

$$\text{var}(\hat{\tau} | s') = \frac{1-c}{f^2 c} \sum_j \frac{n'_{.j}}{n'_{.j} - 1} \cdot \left[\sum_k (n'_{kj} - 1) s'_{kj}{}^2 + \sum_k n'_{kj} (\bar{y}'_{kj} - \bar{y}'_{.j})^2 \right]. \quad (14)$$

Our second and third approximations are to approximate $n'_{.j} / (n'_{.j} - 1)$ by one (A2) and $(n'_{kj} - 1)$ by n'_{kj} (A3) in equation (14). We now require the expectation of the first term in (14) and find

$$E \left[\frac{1-c}{f^2 c} \sum_j \sum_k n'_{kj} s'_{kj}{}^2 \right] \approx \frac{1-c}{f c} \sum_j \sum_k N_{kj} S_{kj}^2. \quad (15)$$

In (15), one further approximation (A4) is necessary; we ignore the possibility of $n'_{kj} \leq 1$ for any k, j . We also require the expectation of the second term in (14). The exact formula turns out to be

$$\frac{1-c}{fc} \sum_j \left\{ \sum_k N_{kj} (\bar{Y}_{kj} - \bar{Y}_{\cdot j})^2 + a_1 \sum_k S_{kj}^2 - a_2 \right\} \quad (16)$$

where $a_1 = 1 - f - E[n'_{kj}(1 - n'_{kj}/N_{kj})/n'_{\cdot j}]$ and $a_2 = E[(\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j}))^2/n'_{\cdot j}]$. We note first that $|a_1| \leq 1$ and thus when combined with N_{kj} in (15), it can be ignored (approximation A5). Also, if in $a_2 n'_{\cdot j}$ is approximated (A6) by its expectation, $fN_{\cdot j}$, since $E[\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j})] = 0$, we have

$$a_2 \approx \frac{1}{fN_{\cdot j}} \text{var} \left(\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j}) \right) \approx (1-f) \sum_k \frac{N_{kj}}{N_{\cdot j}} (1 - W_{kj})(\bar{Y}_{kj} - \bar{Y}_{\cdot j})^2$$

where we have finally approximated $(N_{kj} - 1)$ by N_{kj} . (A7) in computing the variance of the hypergeometric variable n'_{kj} . When compared to the similar term with coefficient N_{kj} in (16), we discover that a_2 itself is approximately negligible. Finally, once again ignoring differences between N_{kj} and $(N_{kj} - 1)$ or between $N_{\cdot j}$ and $(N_{\cdot j} - 1)$ (approximation A8), (15) and (16) can be combined to yield

$$\begin{aligned} E(\text{var}(\hat{\tau} | s')) &\approx \frac{1-c}{fc} \sum_j \frac{N_{\cdot j}}{N_{\cdot j} - 1} \sum_k [(N_{kj} - 1)S_{kj}^2 + N_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j})^2] \\ &= \frac{1-c}{fc} \sum_j N_{\cdot j} S_{\cdot j}^2. \end{aligned} \quad (17)$$

Combining (13) and (17), we finally obtain

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_k N_k S_k^2 + \frac{1-c}{fc} \sum_j N_{\cdot j} S_{\cdot j}^2. \quad (18)$$

The validity of this approximation rests on three facts. First, when (18) is evaluated in the five examples for which simulated data exists, the results compare very favorably. The approximated standard error given by (12) is 113.25, 108.97, 108.09, 106.77, and 106.32 for $f' = .10, .20, .30, .40$, and $.50$, respectively. These values are nearly equal to those in Table 3 and the column giving $SE(\hat{\tau})$ and $M = 0$ with m equal to 500 or 2500. Second, the error introduced by each approximation made was analyzed and found, with the possible exception of approximation A6, to be negligible in the case of relatively large population and sample sizes. Even in the case of A6, the law of large numbers indicates that $n'_{\cdot j}$ will be well approximated by its expectation if the sample sizes are reasonably large. Finally, as described in the following, this approximation formula reduces to the exact formula in all three standard situations. First, this situation reduces to the usual stratified sampling according to pre-strata when we take $J = K$, $P_{\cdot j} = P_k$ for $j = k$, and $c = 1$. Here, formula (18) reduces to $\text{var}(\hat{\tau}) \approx (1-f)/f \sum_k N_k S_k^2$, which is well known to be the exact formula. Also, the estimation scheme described reduces to the usual two phase sampling for stratification when we take $K = 1$ and (18) again reduces to the exact formula (see Cochran 1977, p. 329). Similarly, we obtain the situation of regular stratified sampling by post-strata if we take $f = 1$ (here, K and the pre-stratification become irrelevant), and formula (18) again reduces to the exact value.

REFERENCES

- ALLEN, J.R. (1984). Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS) in the United States. *Seminars in Oncology*, 11, 4-11.
- CASAREALE, D. *et al.* (1984/5). Prevalence of AIDS-associated retrovirus and antibodies among male homosexuals at risk for AIDS in Greenwich Village. *AIDS Research*, 1, 407-421.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.
- HAN, C. (1973). Double sampling with partial information on auxiliary variables. *Journal of the American Statistical Association*, 68, 914-918.
- HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- KURITSKY, J.N. *et al.* (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion*, 26, 205-207.
- LUI, K. *et al.* (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immuno-deficiency syndrome. *Proceedings of the National Academy of Sciences*, 83, 3051-3055.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- REDFIELD, R.R., and BURKE, D.S. (1987). Shadow on the land: the epidemiology of HIV infection. *Viral Immunology*, 1, 69-81.
- ROBERT-GUROFF, M. (1986). Prevalence of antibodies to HTLV-I, -II, and -III in intravenous drug abusers from an AIDS endemic region. *Journal of the American Medical Association*, 255, 3133-3137.
- VARDEMAN, S., and MEEDEN, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.
- WHITE, D. (1987). Mean squared error of estimators using two stage sampling for stratification and prior information. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Sample Design of the 1988 National Farm Survey

C. JULIEN and F. MARANDA¹

ABSTRACT

The National Farm Survey is a sample survey which produces annual estimates on a variety of subjects related to agriculture in Canada. The 1988 survey was conducted using a new sample design. This design involved multiple sampling frames and multivariate sampling techniques different from those of the previous design. This article first describes the strategy and methods used to develop the new sample design, then gives details on factors affecting the precision of the estimates. Finally, the performance of the new design is assessed using the 1988 survey results.

KEY WORDS: Multi-purpose sampling; Multiple frame; Area frame; Multivariate stratification.

1. INTRODUCTION

The National Farm Survey (NFS) is a probability-based sample survey focussing on several subjects related to agriculture in Canada. It is conducted annually in June and July in all provinces except Newfoundland, where a separate survey is carried out.

The previous NFS sample design, dating from 1983, was based on the results of the 1981 Census of Agriculture. A description of it may be found in Ingram and Davidson (1983). However, since 1981 the farm population has changed significantly, reducing the effectiveness of this design. Furthermore, the requirements of the survey have changed somewhat over the years, resulting in the need to update the samples.

A new sample design was therefore developed based on the results of the 1986 Census of Agriculture, and became operational in the summer of 1988.

2. OBJECTIVES OF THE SURVEY

The primary objective of the survey is to provide timely, reliable estimates of levels and annual trends for over 100 agriculture variables. Essentially, these variables may be divided into three categories: cropland areas for the current year; livestock numbers on July 1; and receipts and operating expenses for the previous calendar year. In terms of reliability, the objective of the survey is to obtain coefficients of variation (CV) below 5% at the provincial level for the major parameters.

Survey data are normally summarized to the provincial level. However, primarily for analysis purposes, results for sub-provincial regions are also produced using domain estimation methods.

Another important objective of the survey is to obtain a master sample from which sub-samples are chosen for use in other farm surveys conducted by Statistics Canada.

¹ C. Julien is a methodologist with the Census Data Quality and Analysis Section, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6; F. Maranda is chief of the Agriculture Survey Methods Section, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

3. TARGET POPULATION AND SURVEY POPULATION

The target population includes all farms in the provinces surveyed which received \$250 or more from the sale of agricultural products during the 12 months preceding the survey. Also included are farms which do not meet the \$250 criterion at the time of the survey, but which expect to earn at least this sum during the 12 months following the survey. Such farms, which either began operating just prior to the survey or are temporarily inactive, are relatively few in number.

The survey population, or the group from which the sample is selected, excludes farms operated by institutions as well as those located on Indian reserves or settlements. The terms institution, Indian reserve and Indian settlement are defined in Statistics Canada (1987, pp. 115-117, 145, 152). The cost-benefit ratio associated with collecting data on these types of farms is very high. Because of this, they are excluded in order to enable more efficient use of the resources available for the survey. The contribution of such exclusions to national agricultural production is small and is estimated using adjustment factors which are based on Census data.

4. SAMPLING FRAMES AND THEIR USE

In theory, the survey population is divided into two groups, the first of which includes the farms enumerated in the Census and the second all other farms. These include the undercoverage from the Census and so-called new farms, that is, those which began operating after the Census.

The first group is covered all or in part, depending on the province, by one or two list frames created from the list of census farms. To complement the list frames and ensure complete coverage of the survey population, an area frame, created from the agricultural enumeration areas (EAs), is used. An enumeration area is the geographical region enumerated by a census representative. Furthermore, an EA is said to be agricultural if it contains at least one census farm. An area frame is needed to compensate for the shortcomings of the list frames, particularly their difficulty to identify new farms.

The estimation requirements of the survey and the characteristics of agriculture in Canada vary by region. To better account for these variations, the territory covered by the survey is divided into three regions and a different sample design is used in each one. The three regions involved are: the Prairie provinces and the Peace River district in British Columbia; Quebec and Ontario; and, finally, the Maritime provinces and the rest of British Columbia. The first of these regions is called the Canadian Wheat Board (CWB) region, since the entire region comes under the jurisdiction of this organization.

The total sample size in each of the three regions is essentially based on the overall budget available for data collection. Within each region, sample allocation among the various provinces and, where applicable, among the various frames, depends on several factors. The primary ones are the square root rule applied to the size of the survey population, historical allocations in the survey, and the results of various analyses centred on the expected precision of the estimates.

4.1 The Canadian Wheat Board Region

In this part of Canada, two list frames and one area frame are used in each province.

The first list frame (L1 list) essentially includes the large and medium-sized census farms in relation to key crop, livestock and expense variables. This list is obtained using an iterative process which consists in establishing a threshold for each key variable and including in the

list all farms that exceed at least one of these thresholds. Each threshold is adjusted separately upward or downward so that the L1 list, once completed, includes approximately 35% of the survey population's farms and accounts for 50% to 90% of the total agricultural activity, depending on the key variable in question. These percentages are used because experience has shown that the resulting list is composed of farms which, individually, are more stable over time than the rest of the farms in the survey population. This stability leads to the creation of strata which remain homogeneous over the years, which is a factor in maintaining the efficiency of the sample design.

In each province, the L1 list is then stratified within sub-provincial regions based on nine key variables. A sample of farms is selected and used to obtain data on crops and livestock. Because data on expenses are more difficult and costly to collect, only a sub-sample, called the core sample, is used to obtain this information.

The second list frame (L2 list) includes all census farms with more than 20 acres of cropland which were not included in the L1 list. The L2 list is stratified within crop districts based on a single key variable, namely, cropland area at the time of the Census. The L2 list is used to complement the L1 list for preliminary crop data. These data must be collected within very tight deadlines which, for operational reasons, cannot be met using the area frame.

The area frame includes all agricultural enumeration areas, except those on Indian reserves and in the so-called marginal agricultural regions, that is regions with little agricultural activity. Marginal regions are found mostly in the northern parts of the provinces and in urban fringes. The few census farms located in marginal regions are added to the L1 list, since it is the only list used to collect data on all survey variables.

The area frame is stratified using the same sub-provincial regions and key variables as the L1 list. It ultimately produces a sample of segments which are delineated on topographic maps. The identity of the farmers operating land in one of these segments is obtained through on-site enumeration. Manual matching of names and addresses then enables detection of segment farms overlapping one of the list frames. This detection is essential because each time the area frame is used to complement a list frame, only those segment farms that do not overlap the list in question are used, thus ensuring that the list and area frames represent mutually exclusive domains.

Complete information is required on all segment farms except those overlapping the L1 list, as the data for this list are obtained from the sample selected from it.

4.2 Quebec and Ontario

In each of these provinces, a single list frame, called L1, and an area frame are used.

The list frame is composed of all census farms in the survey population. The methodology used in sampling from this list is similar to that used for the CWB region L1 list, apart from two differences. First, incorporated farms, or farms founded as business corporations, are separated from the other farms, and strata are created independently within the two groups. This preliminary separation is performed because only incorporated farms are required to report their expenses in the survey, since the expenses of the non-incorporated farms are obtained from Revenue Canada tax records. It should be noted that the confidentiality of these records is completely protected under the Statistics Act. Second, sub-sampling for expenses is unnecessary because less than 25% of the farms in the survey population are incorporated.

The area frame and its sample design have not been modified following the last Census, due to a lack of resources. Only the marginal regions were updated, resulting in their enlargement.

4.3 Maritime Provinces and the Rest of British Columbia

In each province of this region, the sample design includes only one list frame, again called L1, which is made up of all census farms in the survey population. Given that a list frame tends to deteriorate with time and that there is no area frame to supplement it, it becomes more difficult to completely cover the survey population. However, because of the relatively small number of farms, under 30 000 in these provinces, more complex procedures were implemented to keep the list up-to-date. Notably, farms which were missed in the Census or which began operating following it may be detected through these procedures. Thus, for all practical purposes, the list frame is considered to ensure full coverage of the survey population.

In each province of this region, the list is stratified and a sample of farms is selected using the same approach as in Quebec and Ontario. All the estimates required are produced from this sample.

5. LIST SAMPLING TECHNIQUES

Samples are taken from the list frames using a one stage, stratified sample design where the farms constitute the sampling units. The strategy and methods used to develop this design are essentially the same, regardless of the province and list involved. However, the combination of methods and key variables used may vary from case to case.

The first step consists in identifying the farms with distinct characteristics and in automatically including them in the sample. There are essentially two kinds of these so-called self-representative or take-all farms. The first group includes those with a unique operating structure such as community pastures and multiholding corporations, while the second group contains the farms which clearly stand out from the majority because of their very large contributions to key crop, livestock and expense variables. Due to the skewness (to the right) of the distributions involved, complete enumeration of these farms is an efficient way to reduce sampling variance.

Farms with very large contributions are identified through an intuitively-based rule which produced good results in the previous sample design. This rule, called the sigma-gap rule, is applied separately to each key variable using all farms having a non-zero value for the variable in question. Farms with a sufficiently high contribution to one of the key variables, as determined by this rule, are said to be take-all.

The sigma-gap rule, as adapted to the survey, functions as follows. Given a univariate distribution of points x_i , $i = 1, 2, \dots, N$, $x_i > 0$ for all i , and given σ as its standard deviation, the points are arranged in increasing order $x_1 \leq x_2 \leq \dots \leq x_N$; for the half of the distribution to the right of the median, the distance between each successive pair of points $d_i = x_i - x_{i-1}$ is determined; given i_o , the smallest i for which $d_i \geq \sigma$, all points $i \geq i_o$ correspond to take-all farms. If $d_i < \sigma$ for all i , no point in this distribution distinguishes itself sufficiently from the others to be declared a take-all farm.

The second step consists in dividing the rest of the farms in the list into take-some strata. In most cases, the strata are formed within sub-provincial regions according to nine key variables representing the usual three categories: crops, livestock and operating expenses. The number of variables in each category is one, six and two respectively.

The underlying principle to the stratification is as follows. Each farm is characterized by nine variables, and neighbouring farms, defined in terms of Euclidian distance, are grouped together. Two multivariate clustering algorithms are used for this purpose. These algorithms are called FASTCLUS and CLUSTER, since they are available in the procedures of the same name in the SAS statistical analysis software package (version 5).

The FASTCLUS algorithm divides a set of observations into a predetermined number of mutually exclusive clusters. First, the algorithm chooses observations which serve as initial cluster seeds. Each observation is then assigned to the nearest seed, and once this is completed, the cluster seeds are updated by the means of the clusters thus formed. The process is repeated until the changes in the seeds become minimal. The FASTCLUS algorithm is based on work by Hartigan (1975) and MacQueen (1967).

The CLUSTER algorithm groups a set of observations into mutually exclusive clusters in a hierarchical structure. Initially, each observation forms a cluster in itself. Based on a technique inspired by Ward (1963), the two most similar clusters are combined into one, which subsequently replaces them. The process is repeated until only one cluster remains. Massart and Kaufman (1983) provide an introduction to this type of classification. Thus, the set of observations is broken down into as many partitions as there were observations to begin with, and each partition corresponds to a stratification.

These algorithms are used successively as follows. FASTCLUS is used first to group the farms into 250 clusters, which are then progressively combined to form the strata using CLUSTER. Initial classification is performed with FASTCLUS, since using CLUSTER directly with a high number of records would require excessive computer time.

Each of the three categories of variables must contribute equally to strata formation. To ensure this, the initial stratification variables are transformed so that the sum of the transformed variables in each category has a mean 0 and a predetermined variance, usually 1. The crop category with its single variable may be standardized in the usual manner by subtracting its mean and dividing by its standard deviation. In each of the other two categories, two successive transformations are performed independently. Given X_i , the initial variables of a given category C, a principal components analysis was performed to obtain transformed variables Y_i . These new variables, with mean μ_i and variance σ_i^2 , are linear combinations of the former ones and mutually independent. The Y_i are then standardized to obtain final stratification variables Z_i as follows:

$$Z_i = \frac{Y_i - \mu_i}{\left(\sum_{i \in C} \sigma_i^2\right)^{1/2}}.$$

Thus, the mean and variance for $\sum_{i \in C} Z_i$ are 0 and 1 respectively.

An empirical approach is used to determine the number of strata. Several stratifications and allocations are performed by varying the number of strata. Then, the coefficient of variation curve is drawn as a function of the number of strata for all key variables and many others. These curves generally resemble Figure 1. Stratification gains are considered to have been virtually fully attained at the point where the majority of curves are practically horizontal. The number of strata chosen is a compromise between this point and the desire to avoid forming too many strata so as to attenuate the effects of incorrect initial classification and stratum jumpers over time, two major causes of outliers or influential observations.

Sample allocation is multivariate and is generally carried out using the same key variables used for stratification. The allocation algorithm consists in minimizing a linear combination of the square of the coefficients of variation of the key variables, within the constraint of a fixed total sample size. Given c_i , coefficient of variation for a key variable, $a_i > 0$ as constant and n_o total sample size, $\sum a_i c_i^2 = f(n)$ must be minimized within the constraint $n = n_o$. The algorithm used is described in Bethel (1986). Adjustments are then made to obtain a minimum sample size of 4 and a maximum weighting factor of 50 in each stratum.

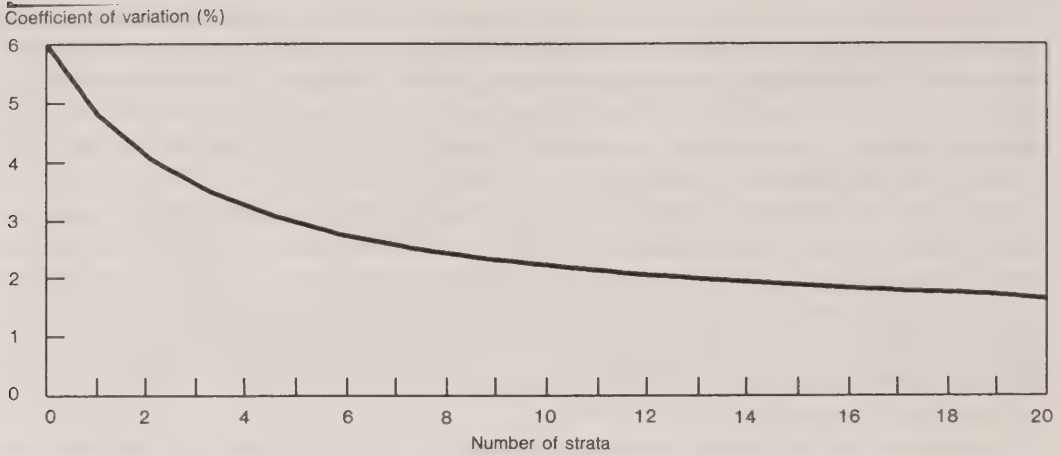


Figure 1. General Curve of the Coefficient of Variation as a Function of the Number of Strata

Finally, once allocation has been completed, the farms are sorted within each stratum by sub-provincial region and total operating expenses and a sample is selected using circular systematic sampling. For the L1 list in the CWB region, the complete sample is chosen first; the core sub-sample is then selected from it using circular systematic sampling.

6. AREA SAMPLING TECHNIQUES

Area samples are selected according to a two-stage stratified sample design. The Census enumeration areas and segments represent the primary and secondary sampling units respectively.

Given that the area sample design has not been modified for Quebec and Ontario, the following paragraphs apply only to the CWB region.

The first step consists in measuring the agricultural activity in each of the frame’s EAs by summarizing to the EA level the data for the census farms not included on the L1 list. Excluding the L1 list farms from the summarization process produces EA distributions which accurately reflect the characteristics of small farms. Subsequent use of these distributions enables an area sample complementing the L1 list with respect to small farms to be selected with greater efficiency.

Once the summarization process has been completed, each EA is treated as a farm for sampling purposes. The EA selection strategy and methods are very similar to those applied

to the CWB region L1 list. First, take-all EAs are determined using the sigma-gap rule. The remaining EAs are then allocated to take-some strata within sub-provincial regions using the CLUSTER multivariate clustering algorithm. Preliminary classification with FASTCLUS is unnecessary in this case due to the relatively low number of EAs, never more than 3000 per province, to be processed. Furthermore, the usual standardizations suffice for transforming the key variables. A principal components analysis was not used because the area frame's contribution to provincial estimates does not justify such an approach.

Allocation to strata is performed with the same algorithm used for the list, and the minimum sample size is again established at 4. The sample size is then divided by four in each stratum, and four separate replicates are selected using circular systematic sampling. Replicates facilitate variance calculation, as a single secondary unit is often chosen per primary unit.

Once the EAs have been selected, their boundaries are traced on topographic maps and they are divided into segments of approximately 7.5 km² (3 mi²). Natural boundaries such as roads and rivers are used as much as possible to facilitate the work of field interviewers. Simple random sampling without replacement of the segments is performed at a minimum rate of 1 out of 30 in each selected EA. There are, however, some exceptions to the rule: additional segments are taken so that the overall weighting factor does not exceed 180; a minimum of two segments are selected in each EA belonging to the strata subjected to first-stage complete enumeration; and, finally, when the same EA appears in more than one replicate, measures are taken to avoid selecting the same segment more than once. Nevertheless, these exceptions are rare.

7. RESULTS OF THE SAMPLE DESIGN

Table 1 contains the results of the list frame sample design. The following items are included: the number of farms in the list (*N*); the number of strata (*H*); the number of farms in the sample (*n*); and, finally, the number of farms in the core sub-sample (*n*-core) in those provinces where it applies.

Table 1
Results of the List Frame Sample Design

Province	L1 List				L2 List		
	<i>N</i>	<i>H</i>	<i>n</i>	<i>n</i> -core	<i>N</i>	<i>H</i>	<i>n</i>
P.E.I.	2,830	26	451				
N.S.	4,273	35	550				
N.B.	3,544	39	498				
Quebec	41,380	80	6,096				
Ontario	72,598	78	8,401				
Manitoba	6,712	48	1,364	490	18,058	29	2,267
Saskatchewan	15,668	48	3,625	1,106	45,798	41	4,573
Alberta	13,928	63	2,981	909	38,504	25	2,973
B.C. (Peace) ^a	494	25	190	190	1,187	6	170
B.C. (rest) ^b	17,042	41	1,999				
Total	178,469	479	26,155	2,695	103,547	101	9,983

^a Peace River district in British Columbia.
^b British Columbia minus the Peace River district.

Table 2 contains the results of the area sample design in those provinces where such a design is used. The following items are indicated: the number of EAs in the frame (N); the number of strata (H); the total number of EAs sampled (n); the number of EAs sampled where each EA is counted only once when it appears in more than one replicate (n -once); and, finally, the number of segments chosen (m).

8. FACTORS AFFECTING THE PRECISION OF THE ESTIMATES

To better appreciate the results obtained from the 1988 survey, three factors affecting the reliability of the estimates must be discussed. These factors are the sample size, the treatment of the total non-response and the estimation methodology.

First, the sample size for the L1 list in the CWB region was reduced by 10% in relation to that of the corresponding list used in the previous sample design. This reduction was prompted mainly by the desire to lower costs.

Second, the methodology used to treat total non-response was modified in 1988. Previously, when a farm failed to respond to the survey, its data were imputed using the data from another farm in the same stratum. These imputed data enabled the sample to be completed to its original size. However, in 1988, the cases of total non-response were not imputed; instead only the respondent sample was used and the weighting factors adjusted upward. The actual sample is therefore reduced in relation to the former method.

In the 1988 survey, the total non-response rate varied between 2% and 13%, depending on the province. The national rate was 10%. Non-response rates are presented in detail in Table 3.

Table 2
Results of the Area Sample Design

Province	N	H	n	n -once	m
Quebec	2,065	43	191	182	230
Ontario	2,687	49	195	185	259
Manitoba	794	21	277	264	305
Saskatchewan	1,496	26	328	308	477
Alberta	1,623	32	328	319	434
B.C. (Peace) ^a	54	7	36	32	58
Total	8,719	178	1,355	1,290	1,763

^a Peace River district in British Columbia.

Table 3
Total Non-response Rate (%) by Province

Province	Refusals	No Contact	Total
P.E.I.	0.00	3.55	3.55
N.S.	0.00	2.18	2.18
N.B.	0.00	1.61	1.61
Quebec	1.71	6.56	8.27
Ontario	2.27	11.11	13.38
Manitoba	3.45	4.03	7.48
Saskatchewan	4.06	6.46	10.52
Alberta	2.68	7.95	10.63
B.C.	1.78	10.28	12.06
Total	2.32	8.11	10.43

The last factor to be discussed is the estimation methodology. The usual estimators corresponding to a stratified simple random sample are used for list frames. For area frames, an estimator described in Wolter (1986 pp. 19-26) and corresponding to a sample design with independent replicates is used. Provincial estimates are obtained by adding the contribution of the list and area frames since, as previously mentioned, these two frames are independent and represent mutually exclusive domains. Details on the estimation methodology are found in Lynch (1988).

9. ASSESSING THE PERFORMANCE OF THE NEW DESIGN

To assess the performance of the new design, the precision of the estimates obtained in 1988 is compared first to that of the 1987 survey, then to the precision anticipated during the development of the sample design.

9.1 The 1988 and 1987 Surveys Compared

Two opposite tendencies are in effect in a comparison of the precision of the estimate in the 1988 and 1987 surveys. The 1988 estimates should be more precise because the 1987 sample design was already four years old. However, the two sample size reduction factors described in section 8 would indicate less precise estimates for 1988.

Precision is compared using the coefficient of variation of the provincial estimates obtained by combining the L1 list and area frames. The estimates used are those for several key variables whose coefficient of variation in 1987 did not exceed 20%.

The precision of 234 estimates is compared in the charts in Figure 2, where each square represents the CV achieved in 1987 on the x-axis and achieved in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X/2$, $Y = X$ and $Y = 2X$ is also presented.

Nearly 60% of crop estimates were more precise in 1988 than in 1987. The majority of those that were less precise were so to a small degree only. Close to 95% of livestock estimates were more precise in 1988 than the previous year; in fact, 32% of the estimates were even twice as precise. Finally, over 60% of operating expense estimates were more precise in 1988. Some of the 1987 estimates were a good deal less precise, and 7% were even two times less precise. The latter are from Quebec and Ontario, where data on operating expenses are collected from incorporated farms only. Further more, the legal status of a farm in these provinces is difficult to identify, both in the Census and the survey.

Despite the reduction in the effective sample due to total non-response and cutbacks during the sample design development stage, the 1988 survey generally provided more precise estimates for each category of variables.

9.2 Precision Obtained Versus Precision Anticipated

The precision obtained is expected to be inferior to the precision anticipated for two reasons. First, when the weighting factors are adjusted to account for the total non-response, the variance increases slightly. Second, the data used to create the sampling frame were taken from the 1986 Census of Agriculture. These data are subject to error and the sampling frame deteriorates with changes in agricultural activity.

Precision is compared using the coefficient of variation of L1 list frame provincial estimates only. These estimates are for several key variables whose anticipated CV did not exceed 20%.

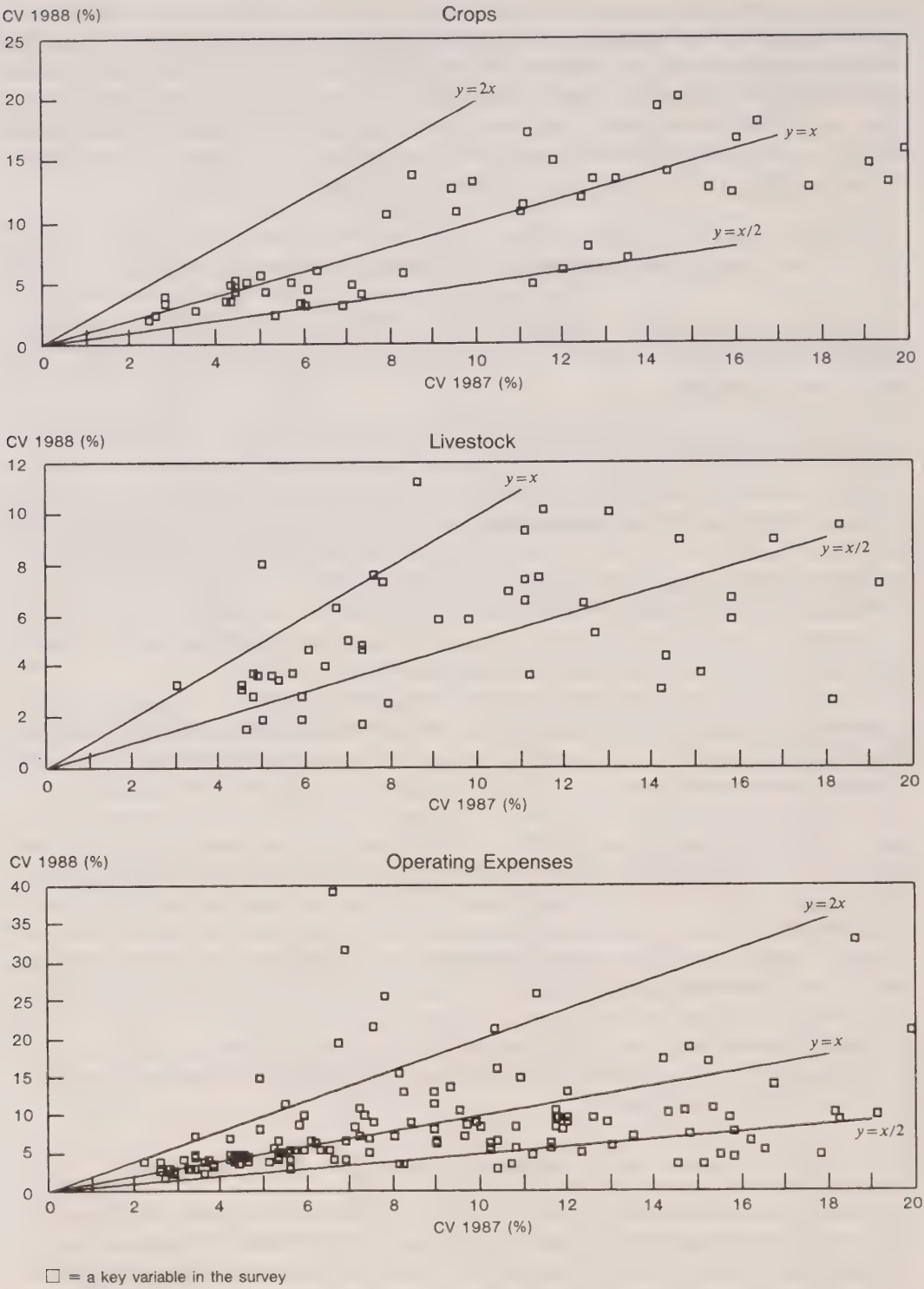


Figure 2. Comparison of the Precision of Key Variable Estimates in the 1987 and 1988 Surveys by Category of Questions.

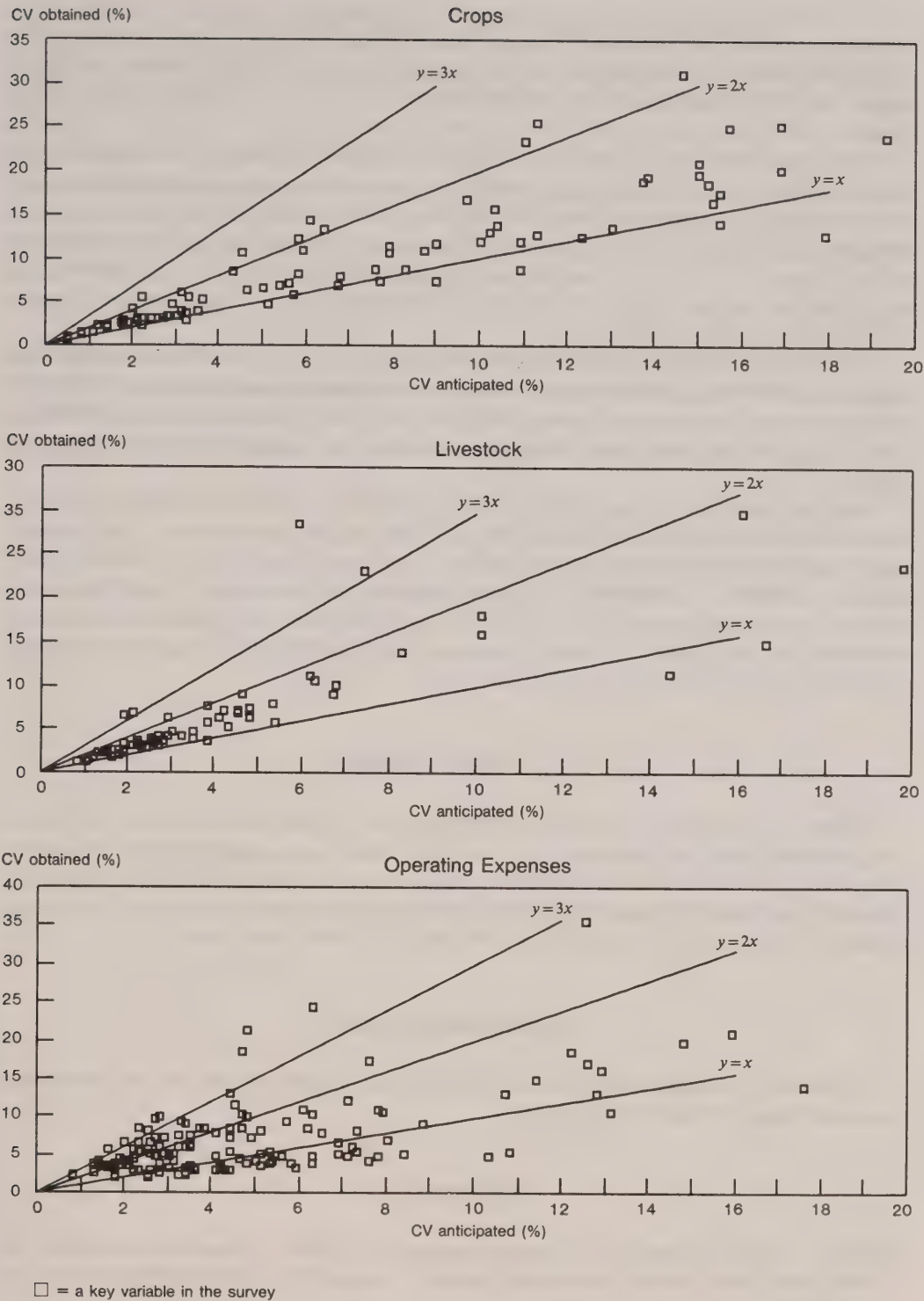


Figure 3. Comparison of the Precision of Key Variable Estimates Obtained in the 1988 Survey and the Precision Anticipated during Development of the Sample Design.

A comparison of the precision of 288 estimates is presented in chart form in Figure 3. In these charts, each square represents the anticipated CV on the x-axis and the obtained CV in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X$, $Y = 2X$ and $Y = 3X$ is shown in the charts.

For the crop and livestock categories, approximately 90% of the estimates are sufficiently precise, given the non-response rate, as most of the key variables are located closer to straight line $Y = X$ than to straight line $Y = 2X$. Two tendencies can be seen for the operating expense estimates. Surprisingly, the CV obtained is lower than the anticipated CV in 28% of the cases, the vast majority of which are found in the CWB region. However, 31% of all estimates are more than two times less precise than anticipated. These cases are found in Quebec and Ontario for the reasons given in section 9.1.

Finally, a complementary study was conducted in which the precision obtained was compared to the anticipated precision based on the size of the sample actually observed. This study revealed that the frequency of estimates at least two times less precise than anticipated dropped from 12% to 5% for crops, from 9% to 5% for livestock and from 31% to 7% for operating expenses.

These studies show that in general the precision obtained is acceptable and differs from the anticipated precision mainly because of the treatment for total non-response. This indicates that the sample design is therefore sound and the L1 list frame is adequate. On the other hand, less precise estimates were obtained for operating expenses due to a problem in identifying incorporated farms in Quebec and Ontario in the Census and in the survey. Finally, the list frame, which was two years old at the time of the survey, was observed to have deteriorated somewhat due mostly to bankruptcies and farm sales.

10. CONCLUSION

In general, survey results were substantially improved following implementation of the new sample design. Moreover, the reduction in sample sizes led to cost savings and a considerable reduction in the response burden on the farmers surveyed. Difficulties remain, however, especially regarding the operating expense variables for incorporated farms in Quebec and Ontario. Further studies to resolve these difficulties are being envisaged.

ACKNOWLEDGEMENTS

The authors would like to thank the editor of the journal and the referees, whose valuable comments helped to improve this article.

REFERENCES

- BETHEL, J. (1986). An optimum allocation algorithm for multivariate surveys. Technical report of the United States Department of Agriculture, Statistical Reporting Service, Statistical Research Division, No. SF and SRB-89.
- GERMAIN, M.-F., DOLSON, D., and MARANDA, F. (1989). Redesign of the 1988 National Farm Survey. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.

- HARTIGAN, J.A. (1975). *Clustering Algorithms*. New York: John Wiley and Sons.
- INGRAM, S., and DAVIDSON, G. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 220-225.
- LYNCH, J. (1988). Cas spéciaux d'estimation dans l'enquête nationale des fermes de 1988. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.
- MacQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability*, 1, 281-297.
- MASSART, D.L., and KAUFMAN, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley and Sons.
- SAS Institute Inc. (1985). *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS institute.
- STATISTICS CANADA (1987). 1986 Census Dictionary. Catalogue 99-101E, Statistics Canada.
- WARD, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Does the Method Matter on Sensitive Survey Topics?

DAVID A. HAY¹

ABSTRACT

The effects of utilizing a self-administered questionnaire or a personal interview procedure on the responses of an adolescent sample on their alcohol consumption and related behaviors are examined. The results are generally supportive of previous studies on the relationship between the method of data collection and the distribution of responses with sensitive or non-normative content. Although of significance in a statistical sense, many of the differences are not of sufficient magnitude to be considered significant in a substantive sense.

KEY WORDS: Data collection; Personal interview; Self-administered questionnaire; Response errors; Alcohol consumption.

1. INTRODUCTION

To "questionnaire" or to interview that is the question to be answered by researchers in the design and conduct of sample surveys on delicate or sensitive topics. The decision on whether to utilize personal or telephone interviews or a variant of the self-administered questionnaires, or a combination thereof, is a critical decision that survey researchers have to make in attempting to optimize the quality of the resultant data.

Encompassed by the more general problems of reliability and validity associated with self-reports of attitudes, behaviour and other phenomena of interest to survey practitioners, is the question regarding the relative merits of the interview and self-administered formats in minimizing or reducing non-sampling biases or errors. In other words, then would different results be obtained from the utilization of different modes of data collection (Smith 1975)?

As far back as 1959, Selltiz *et al.* (1959) stated that most questionnaires and interviews were utilized without evidence of their relative merits. More recently, this position has been re-emphasized by Knudsen *et al.* (1967), Alwin (1977) and Newton *et al.* (1982) who maintain that the selection of the survey mode to be utilized is based on convenience, relative costs and other practical considerations rather than on their methodological adequacy and potential response effects. The planning of survey research, Newton *et al.* emphasize should be determined by what is reliably known about the relationship between methods of administration and response patterns, rather than just on the issues of relative costs, respondent motivation and other similar considerations.

Some studies which have compared personal interviews with more anonymous formats such as self-administered questionnaires or telephone interviews have found minimal and/or statistically non-significant differences in the responses to a variety of topics including those of a private or sensitive nature (DeLameter and MacCorquodale 1975; Gibson and Hawkins 1968; Krohn *et al.* 1974; McDonagh and Rosenblum 1965; Metzner and Mann 1952, Newton *et al.* 1982 and Sykes and Collins 1987.) Other researchers have observed that more candid, self-revelatory and informative responses are more likely to be made by questionnaire and telephone respondents than personal interviewees on topics concerning deviant, sensitive

¹ David A. Hay, Associate Professor, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 0W0.

or embarrassing behaviours and attitudes. (Cannell and Fowler 1963; Ellis 1947; Hubbard *et al.* 1976; Knudsen *et al.* 1967; Siemiatycki 1979; Whitehead and Smart 1972 and Wiseman 1972).

The conclusions of the latter studies were generally based on the untested assumption that the increased reporting of deviant, threatening or embarrassing information was more accurate (Blair *et al.* 1977). This point was also emphasized by Schuman (1980) who stated that frequently no external validation data were obtained, but the researchers "assumed that the more such behaviour was reported, the more accurate the reports – a plausible but not air-tight assumption for most of the topics they dealt with."

The present note is concerned with a further comparison of the relationship between personal interviews and self-administered questionnaires and responses obtained from an adolescent population on a "threatening" or deviant topic, namely alcohol consumption. The results being reported are based on a secondary analysis of data from a study of alcohol-related attitudes and behaviors from a sample of teenagers in a Western Canadian province completed in 1977-78 (Hetherington *et al.* 1978 and 1979).

The study which utilized both personal interviews and self-administered questionnaires provides a unique opportunity to compare the potential effects of the mode of data collection on the resultant data. This type of comparison of interest to survey practitioners is generally not possible in the majority of surveys which tend to rely on one method of data collection.

A stratified random sample of 1502 students in grades 6 to 12 was selected from three school regions in the Province of concern. The total sample of students was randomly assigned by grade to either the self-administered questionnaire or to the personal interview procedure. Approximately one half of the students from each grade 6 to 12 were thus allocated to one of the procedures. The number of students assigned to be interviewed was 752 with 750 students being assigned to the questionnaire data collection.

The questionnaire was group administered by a trained researcher in a room made available at each school for that purpose. The interviews were conducted by fifteen interviewers specifically trained for the study.

The survey instrument which consisted of 75 questions was identical in content for both the interview and questionnaire data collection procedures. The majority of the questions were closed ended and required an average of 20 minutes for completion in both types of administration.

2. RESULTS AND DISCUSSION

A comparison of the personal interview and self-administered questionnaire respondents on a number of personal and familial characteristics was conducted to determine if the two groups differed in respects other than the method of data collection. The results indicated that the two groups did not differ by more than could be attributed to chance on variables such as sex, age, grade of enrollment, parent's educational and occupational backgrounds and religious affiliation. A statistically significant difference was observed on the variable of ethnicity with a higher percentage of Canadian identities reported by the interview respondents.

With the exception of ethnic background, the subsequent analysis was, therefore, based on the assumption that the interview and questionnaire respondents were equivalent on a number of variables that could potentially confound the comparison of obtained responses to the two procedures.

Table 1
Frequency Distribution and Z Probabilities on Selected Questions
for Interview and Questionnaire Results

Variable	Interview (<i>n</i> = 752)	Questionnaire (<i>n</i> = 750)	Two-tailed Z Probability
Ever drink	62.63	73.73	.000
Ever used cigarettes	29.78	37.60	.001

2.1 Variable Distribution

A comparison of the mean responses or frequency distributions for the interview and questionnaire respondents on a number of questions with non-normative or illegal content lent general support to previous research on similar issues. The questions of primary concern are those related to the consumption of alcohol which are viewed as possessing a considerable degree of threat or deviant content for the population under consideration, the majority (99.8%) of whom were under the legal drinking age at the time of the study.

The frequency distributions in Table 1 indicated that a significantly higher percentage of the questionnaire respondents reported ever having more than a sip or taste of an alcoholic beverage. Similar statistically significant differentials were observed between the interview and questionnaire respondents on reported smoking.

For those respondents reporting that they had consumed a drink of alcohol, the mean drinking levels and average age at first drink shown in Table 2 were also suggestive that the questionnaire respondents are more likely to report on deviant behaviour than were their interview contemporaries. The significantly higher average drinking levels for the questionnaire respondents reflects their reporting higher amounts and frequencies of alcohol consumption. The significantly higher average age at first drink for the interviewees indicates their reporting taking their first substantial drink at an older age than did the questionnaire respondents.

Significant differentials between the interview and questionnaire respondents were also observed on the reporting of parental drinking and on the importance of religion in the home questions. The mean values for these three questions indicated that the questionnaire respondents reported higher drinking levels for their parents than did the interviewees and that religion was perceived as being less important in the homes of the questionnaire respondents. While not possessing the same degree of self revelation or threat to the respondent per se, the differentials were viewed as suggestive of an attempt on the part of the interviewees to portray a more favourable or socially acceptable image about their family life.

However, the greater importance of religion in the home reported by the interviewees was not carried through in their self-descriptions of the importance of religion. The statistical equivalence of the means values on the importance of religion to self indicated that the interview respondents were no more likely to report that religion was important to self than were the questionnaire respondents. The two groups of respondents were also equally likely to report on the drinking habits of friends or peers.

The response patterns on other questions possessing somewhat different aspects of ego-involvement or image favourability did not generally support the potential operation of a social desirability effect as was evident for the alcohol related behaviours. As indicated in Table 2, the questionnaire respondents reported receiving significantly higher school grades, had higher educational aspirations in terms of their future educational plans and reported more positive self images on 4 of the 7 self-esteem items and on the composite self-esteem index. Contrary

Table 2
Means, Standard Deviations and "t" probabilities on Selected Questions
for Interview and Questionnaire Respondents

Variable ¹	Interview (<i>n</i> = 752)		Questionnaire (<i>n</i> = 750)		Two-tailed “ <i>t</i> ” Probabilities
	<i>X̄</i>	<i>SD</i>	<i>X̄</i>	<i>SD</i>	
Alcohol and Related Behaviour					
Drinking level	2.31	2.92	2.76	3.05	.003
Age at first drink ^a	3.93	1.32	3.64	1.39	.001
Father drinks	1.82	0.62	1.90	0.58	.011
Mother drinks	1.70	0.50	1.75	0.51	.025
Friends drink	1.92	0.57	1.94	0.56	.481
Educational Variables					
Grades received	4.37	1.49	4.58	1.46	.008
Educational plans	3.02	1.24	3.25	1.24	.001
Religious Variables					
Importance of religion in the home	3.37	1.16	3.15	1.22	.000
Importance of religion to student	3.22	1.12	3.13	1.18	.130
Self-Esteem Indices					
Item 1	2.98	0.60	3.12	0.60	.000
Item 2	2.96	0.49	3.08	0.54	.000
Item 3	3.14	0.55	3.27	0.61	.000
Item 4	2.98	0.51	2.05	0.57	.033
Item 5	3.10	0.63	3.01	0.75	.017
Item 6	2.93	0.56	2.97	0.59	.207
Item 7	3.07	0.54	3.12	0.60	.132
Composite	21.17	2.39	21.65	2.85	.001

^a - Mean value calculated on grouped data.

¹ Variable Codes: Drinking level; composite index of frequency and volume of alcohol consumed 0 = abstainer to 9 = frequent consumer of large amount of alcohol.

Age at first drink: 1 = 6 years or less; 2 = 7-8 years; 3 = 9-10 years; 4 = 11-12 years; 5 = 13-14 years; 6 = 15-16 years; and 7 = ≥ 17 years.

Father, mother and friends drink: 1 = never drinks; 2 = drinks sometimes; 3 = drinks a lot.

Grades received: 1 = mostly D's and F's; 2 = Mostly C's and D's; 3 = mostly C's; 4 = mostly B's and C's; 5 = mostly B's; 6 = mostly A's and B's and 7 = mostly A's.

Educational plans: 1 = will not finish grade 12; 2 = will finish grade 12 only; 3 = will take technical training; 4 = will attend university and 5 = will go to graduate or professional school.

Self-esteem items and index: 1 = strongly disagree; 2 = disagree; 3 = agree and 4 = strongly agree. The additive index for the 7 items ranged from 7 to 28.

to the expectation that the interviewees would attempt to portray a more favourable image, these results tended to indicate that they were more modest in the reporting of school grades received, in their educational aspirations and in their self perceptions. However, the greater anonymity and potential freedom afforded the questionnaire respondents to more willingly report on their alcohol related behaviors may also have resulted in a similar perceived freedom to aggrandize their own merits in relation to these questions on school grades, educational plans and their self conceptions.

However, the presence of a significant distributional response bias between the interview-questionnaire data collections is evident only in the statistical sense of the term. The statistically significant mean value differences on the questions of concern ranged from 0.05 to a maximum of 0.48 on the composite self-esteem index. Given the potential presence of other errors of measurement, the interview-questionnaire response differentials obtained in the present study are not of sufficient magnitude to be considered as indicative of a response bias effect of substantive or practical importance.

Due to the unavailability of reliable information on the actual drinking habits of the students and their parents, the school grades and other responses under consideration, it was not possible to conduct an evaluation of the relative accuracy of the interview and questionnaire responses. As a result it is not possible to indicate the relative superiority of either the self-administered mode or the personal interview for the question responses under consideration. Both types of responses may be subject to an under- or over-reporting bias of an indeterminant direction and/or magnitude.

The results of this note are in general agreement with Bradburn and Sudman (1979) who indicate that no consistent relationship appears to exist between the method of survey administration and the over-reporting of socially desirable behaviour or the under-reporting of socially undesirable behaviors and attitudes. As a result Bradburn and Sudman (1979) and Locander *et al.* (1976) suggest that no data collection procedure is clearly superior for all types of threatening or other questions of concern to survey practitioners.

ACKNOWLEDGEMENT

The author is grateful to S. Parvez Wakil for his critical comments on the original version of this paper and to the anonymous referees and M. P. Singh for their very helpful comments.

The initial study was supported by Health and Welfare Canada Non-Medical Use of Drugs Directorate (Grant #1213-7-10) with additional support from the Applied Research Unit, Psychiatric Research Division, Saskatchewan Department of Health. The study's grantees are also acknowledged for their generosity in allowing the use of the data for this paper.

REFERENCES

- ALWIN, D.F. (1977). Making errors in surveys: an overview. *Sociological Methods and Research*, 6, 131-151.
- BLAIR, E., SUDMAN, S., BRADBURN, N.M., and STOCKING, C. (1977). How to ask questions about drinking and sex: response effects in measuring consumer behavior. *Journal of Marketing Research*, 14, 316-321.
- BRADBURN, N.M., and SUDMAN, S. (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- CANNELL, C.F., and FOWLER, F.J. (1963). Comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly*, 27, 250-264.
- DeLAMETER, J., and MacCORQUODALE, P. (1975). The effects of interview schedule on reported sexual behavior. *Sociological Methods and Research*, 4, 215-236.
- ELLIS, A. (1947). Questionnaire versus interview methods in the study of human love relationships. *American Sociological Review*, 12, 541-553.
- GIBSON, F.W., and HAWKINS, B.W. (1968). Interview versus questionnaires. *American Behavioral Scientist*, 12, 9-11.

- HERZOG, A.R., RODGERS, W., and KULKA, R.A. (1983). Interviewing older adults: a comparison of telephone and face-to-face modalities. *Public Opinion Quarterly*, 47, 405-418.
- HETHERINGTON, R.W., DICKINSON, J., CIPYWNYK, D., and HAY, D.A. (1978). Drinking behavior among Saskatchewan adolescents. *Canadian Journal of Public Health*, 69, 315-324.
- HETHERINGTON, R.W., DICKINSON, J., CIPYWNYK, D., and HAY, D.A. (1979). Attitudes and knowledge about alcohol among Saskatchewan adolescents. *Canadian Journal of Public Health*, 70, 247-259.
- HUBBARD, R.L., ECKERMAN, W.C., and RACHAL, J.V. (1976). Methods of validating self-reports of drug use: a critical review. *Proceeding of the Social Statistics Section, American Statistical Association, Part I*, 406-409.
- KNUDSEN, D., HALLOWELL, D., and IRISH, D.P. (1967). Response differences to questions on sexual standards: an interview-questionnaire comparison. *Public Opinion Quarterly*, 21, 290-297.
- KROHN, M., WALDO, G.P., and CHIRICOS, T.G. (1974). Self-reported delinquency: a comparison of structured interviews and self-administered checklists. *The Journal of Criminal Law and Criminology*, 65, 545-553.
- LOCANDER, W., SUDMAN, S., and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of American Statistical Association*, 71, 269-275.
- MCDONAGH, E.C., and ROSENBLUM, A.L. (1965). A comparison of mailed questionnaires and subsequent structured interviews. *Public Opinion Quarterly*, 29, 131-136.
- METZNER, H., and MANN, F. (1952). A limited comparison of two methods of data collection: the fixed alternative questionnaire and the open-ended interview. *American Sociological Review*, 17, 486-491.
- NEWTON, R.R., PRENSKY, D., and SHUESSLER, K. (1982). Form effect in the measurement of feeling states. *Social Science Research*, 11, 301-317.
- SCHUMAN, H. (1980). Review of improving interview method and questionnaire design. *Social Forces*, 59, 325-326.
- SELLTIZ, C., JAHODA, M., DEUTSCH, M., and COOK, S.W. (1959). *Research Methods in Social Relations* (Revised). New York: Holt, Rinehart and Winston.
- SIEMIATYCKI, J. (1979) A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health*, 69, 238-245.
- SMITH, H.W. (1975). *Strategies of Social Research*. Englewood Cliffs: Prentice Hall.
- SYKES, W.M., and COLLINS, M. (1987). Comparing telephone and face-to-face interviewing in the United Kingdom. *Survey Methodology*, 13, 15-28.
- WHITEHEAD, P.C., and SMART, R.G. (1972). Validity and reliability of self-reported drug use. *Canadian Journal of Criminology and Corrections*, 14, 83-89.
- WISEMAN, F. (1972). Methodological bias in public opinion polls. *Public Opinion Quarterly*, 36, 105-108.

Use of Cluster Analysis for Collapsing Imputation Classes

E.R. LANGLET¹

ABSTRACT

The problem of collapsing the imputation classes defined by a large number of cross-classifications of auxiliary variables is considered. A solution based on cluster analysis to reduce the number of levels of auxiliary variables to a reasonably small number of imputation classes is proposed. The motivation and solution of this general problem are illustrated by the imputation of age in the Hospital Morbidity System where auxiliary variables are sex and diagnosis.

KEY WORDS: Item nonresponse; Auxiliary variables; Imputation matrix; Donors; Disjoint techniques; Hierarchical techniques; Cluster seeds.

1. STATEMENT OF THE PROBLEM

In surveys, the problem of item nonresponse occurs when some but not all information is collected for a sample unit or when some information is deleted because it fails to satisfy edit constraints. In many surveys, this problem is handled by random imputation within classes, a common form of hot deck imputation method. For this type of imputation, a respondent is chosen at random within an imputation class defined by one or more auxiliary variables and the respondent's value is assigned to the nonrespondent.

The problem considered in this paper can be defined as follows. The classifications of the respondents according to certain auxiliary variables form a multi-dimensional imputation matrix where the number of imputation classes equals the number of cross-classification cells defined by the auxiliary variables. If the number of imputation classes is very large, few or no donors may be available in several classes. In addition, manipulation of this large matrix could be very cumbersome computationally. These problems can be alleviated by collapsing the cells of the matrix either by grouping the cells themselves, or the rows, columns or along some other dimension (or combination of dimensions) so that the resulting groups will be homogeneous with respect to the variables requiring imputation. We propose to use cluster analysis to achieve the desired level of collapsing. For this purpose, the values of the variables of interest from donors (or respondents) for each imputation class can be used to assign numerical scores to each class. In this paper, measures based on empirical distribution function for respondent data are used to quantify imputation classes. Cluster analysis can then be used to group the cells of the matrix according to these numerical scores. It will be shown that cluster analysis is appropriate for the problem under consideration. Related useful references concerning the application of cluster analysis to stratify primary sampling units are Drew, Bélanger and Foy (1985), Judkins and Singh (1981) and other references contained therein.

The above mentioned problem arose in the context of age imputation in the Hospital Morbidity System (HMS). This system uses the auxiliary variables sex and diagnosis as the basis for imputing the age. The number of imputation classes were over 5,000 for each sex. A solution based on the technique of cluster analysis was proposed in order to collapse the levels of

¹ E.R. Langlet, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

the diagnosis variable to 40 groups of related diagnoses. In section 2, a brief review of the commonly used cluster analysis techniques is presented. Use of cluster analysis for the problem of collapsing imputation classes is illustrated for the example of imputation of age for the HMS data in section 3 including the relative performance of the proposed method with respect to the current method. Both methods utilize a hot deck approach but the proposed method redefines the imputation classes using cluster analysis. Some concluding remarks including possible generalizations of the method are given in section 4.

2. CLUSTER ANALYSIS TECHNIQUES: A BRIEF REVIEW

The problem of classifying a given number of entities described by a number of quantitative variables into groups such that entities within the same groups or clusters will be similar to each other and dissimilar to entities in different groups is considered in this section. A good review of clustering techniques is given by Everitt (1980) mainly based on the work of Cormack (1971). Most clustering techniques can be classified into two groups, namely 'hierarchical techniques' and 'disjoint techniques', the latter one also known as 'optimization techniques'. These two groups of techniques will be described below. Some other methods, are density techniques where clusters are formed by searching for regions containing dense concentrations of entities. This is based on the fact that if entities are described as points in a metric space, there should be parts of the space in which the points are very dense, separated by parts of low density. Another class of techniques is called clumping techniques in which the clusters can overlap. In certain fields such as language studies, for example, classification must permit an overlap between the classes because words tend to have several meanings, and if they are classified by their meanings they may belong in several places.

Hierarchical techniques can be subdivided into 'fusion techniques' and 'divisive techniques'. In fusion methods, each entity begins in a cluster by itself. At each step, the two closest clusters are fused to form a new cluster until only one cluster containing all the observations is left. In divisive techniques, all entities are first grouped into one cluster. Then, at each step, groups of the entities are successively broken down into finer partitions until each entity constitutes a cluster by itself. Hierarchical techniques differ with respects to the definition of the distance measure between observations or groups of observations. An advantage of hierarchical techniques is that a single run can produce results for one cluster to as many as you like by stopping the fusion or division process at the desired level of the hierarchy. Obviously, hierarchical techniques can be used for only small data sets since there are $n(n - 1)/2$ possibilities to fuse two entities in a group of n entities and $2^{n-1} - 1$ possibilities to break a group of n entities in two groups.

In contrast to hierarchical techniques where observations belong to a series of clusters depending on the level of the hierarchy, disjoint techniques divide observations into a number of clusters (generally predetermined) such that each observation belongs to one and only one cluster. They also differ from hierarchical techniques in that they admit relocation of the observations so that a poor initial partition can be corrected at a later stage. Disjoint techniques are clearly more appropriate than hierarchical techniques to handle large data sets. Disjoint techniques are also called optimization techniques because they seek for a partition of the data which optimizes some predefined criterion. Various disjoint techniques differ in the way the methods obtain an initial partition and in the clustering criterion they try to optimize. Usually, disjoint techniques start by selecting a set of points called cluster seeds as a first guess of the means of the clusters. A number of procedures have been suggested for choosing these points

(Anderberg 1973). Once the cluster seeds have been selected, the entities are then assigned to the closest cluster seeds (usually, the Euclidean distance is used). Estimates of the cluster means might be updated after each allocation (MacQueen 1967) or after all entities have been allocated (Ball and Hall 1967). Once an initial partition has been found (which is equivalent to finding a set of cluster seeds and to allocating each entity to the closest cluster seed), a search is made for entities whose re-allocation to some other group will improve the clustering criterion. This procedure is repeated until no further move of a single entity improves the clustering criterion. A local optimum is then reached. This is what Anderberg (1973) calls 'nearest centroid sorting'. In general, there is no way to know whether a global optimum has been reached.

3. APPLICATION: FORMING IMPUTATION CLASSES FOR THE HMS

3.1 Background

The Hospital Morbidity System (Statistics Canada 1987) consists of a count of inpatient cases, discharged during the year from general and allied special hospitals in Canada except Yukon and Northwest Territories. Each record of the system contains at least one diagnosis code, the age and sex of the patient, the length of stay, *etc.* The first valid diagnosis on the record is called the tabulating diagnosis and is the diagnosis on which tabulations are based in the publications. This diagnosis can be seen as the main cause for which the patient is hospitalized and is coded according to the 9th Edition of the International Classification of Diseases (World Health Organization 1977) which contains more than 5,000 diagnoses.

The age imputation problem in the HMS is currently treated by a hot deck method. In this imputation problem to predict the age of the patient y , two auxiliary variables are used, namely the tabulating diagnosis d which is always present on the record and the sex of the patient s . The sex itself needs to be imputed first if it is missing according to the observed male/female proportions of d over previous years. Classification of the patients according to d and s forms an imputation matrix with the number of imputation classes larger than 5000×2 . In order to reduce the dimension of the imputation matrix, diagnoses were regrouped or collapsed, based on the age distribution of each diagnosis. Let F_d denote the age distribution in the population of the patients with tabulating diagnosis d . Then, diagnoses A and B would be collapsed together if F_A is close to F_B . Estimates of F_d from available data can be used for this purpose. It should be noted that the sex variable was not used in defining imputation classes (see section 4 for details on how it could be used) although it was used in the imputation scheme. By not using the sex variable for defining imputation classes, the number of imputation classes of the imputation matrix is reduced by half.

In order to motivate the proposed method for collapsing imputation classes, we will first describe the current method and its limitations. The collapsed groups were created by comparing manually (using histograms) the shapes of the empirical age frequency distributions, \hat{F}_d of all diagnosis codes corresponding to 1974 HMS data. Thirty six groups were obtained and a 37th group was created for those diagnoses for which less than 200 observations were available. The number of groups was determined a posteriori arbitrarily. The main deficiency of the current method comes from the fact that no statistical criterion was used to group diagnoses which makes the method labour intensive and somewhat subjective. These groups were obtained by simply comparing histograms. An evaluation of the current imputation method indicated that the resulting groups of diagnoses were, in a few cases, not homogeneous with respect to \hat{F}_d and consequently needed to be updated.

3.2 Proposed Method

The proposed method can be briefly described as follows. We shall consider the case when only one quantitative variable needs to be imputed. Extension to cases where more than one variable requires imputation is discussed in section 4. Let's denote by y the variable to be imputed and by F_i the distribution of variable y in class i . Note that the classes are defined by the cross-classification of one or more auxiliary variables which are suitably categorized if necessary. The first step is to find an appropriate set of parameters to represent F_i in each class, for example, the first three or four moments of the F_i 's or the percentiles. The next step is to estimate these parameters from the respondent data. Finally, a suitable technique of cluster analysis on the set of estimated parameters can be used to condense the number of classes such that classes grouped together will be similar with respect to the parameters representing the F_i 's.

A justification for the choice of the proposed method in the context of the age imputation for the Hospital Morbidity System (HMS) will now be presented. First, consider some possible alternative strategies to the collapsing problem. One strategy for this problem might be similar to the original method that was used for 1974 data, that is, to group diagnoses according to the distributions \hat{F}_d but using a statistical criterion for grouping instead of manually comparing histograms. Data would be cross-classified by tabulating diagnoses, sex and a number of age groups, say 10. Two diagnoses would be grouped together if the proportion of cases in each of these ten age groups, p_1, \dots, p_{10} were judged to be close to each other according to some criterion such as the Euclidean distance or a chi-square measure. Note that the use of a chi-square measure would cause serious computational burden since no commonly available cluster analysis program uses this distance measure. This would imply the calculation of the chi-square distance for all possible pairs of diagnoses. Another possible strategy would be to first use data reduction techniques such as principal components to reduce the dimension of age groups and then decide whether two diagnoses are close based on principal component scores. An obvious disadvantage to all these methods is the number of observations required to obtain a reliable estimate of the categorical age distribution for each diagnosis.

In view of the above problem, we decided to use the first two or three moments to approximately describe F_d . We started with three – the mean m_d , the standard deviation s_d and the skewness coefficient b_d . However, it was found by means of principal component analysis that it was not necessary to include b_d . The approach then is to collapse diagnoses according to the sample mean, m_d , and the sample standard deviation s_d . Cluster analysis can be used to provide a suitable statistical technique for this purpose. An obvious advantage with this approach over other strategies based on the categorical distribution of age is that a reliable estimation of two moments requires much fewer observations than the estimation of the proportion of cases over several age groups. In section 4, implementation of this approach is described for the problem of age imputation.

3.3 Procedure Steps in the Implementation of the Proposed Method for HMS Data

There are four steps in implementing the proposed collapsing method based on cluster analysis for the age imputation problem for HMS data.

Step I: Selection of a clustering method

Before selecting a clustering method, it should be noted that our goal is primarily to partition the diagnoses into homogeneous groups without trying to uncover 'natural' or 'real' clusters. This is called 'data dissection' in the literature (Everitt 1980). Another important consideration is the availability of a well tested clustering program using an efficient

clustering method. The determinant consideration for the selection of a clustering method was the number of observations in our data set which resulted in the selection of a disjoint technique rather than a hierarchical technique.

Taking into consideration the above points, the disjoint clustering technique used in the FASTCLUS procedure of SAS (1985) was chosen to do the analysis. This procedure performs a disjoint cluster analysis based on the usual Euclidean distances computed from a given set of quantitative variables. The FASTCLUS procedure combines an effective method for finding initial clusters (or initial clusters can be given by the user) with a standard iterative algorithm for minimizing the sum of squared distances from the cluster means. FASTCLUS was directly inspired by Hartigan's leader algorithm (1975) and MacQueen's k -means algorithm (1967). A set of cluster seeds is first selected as a guess of the means of the clusters. Each observation is assigned to the nearest cluster seed to form temporary clusters. The cluster seeds are replaced by the means of the temporary clusters each time an observation is assigned (this is an option chosen for our application). After each pass through the data set, the observations are assigned to the nearest cluster seed until the changes in the cluster seeds become small or null (chosen to be null for our application). The final clusters are formed by assigning each observation to the nearest cluster seed.

Step II: Estimation of parameters

Two years of HMS data from 82–83 and 83–84 fiscal years were gathered to get estimates m_d and s_d for each diagnosis d . These estimates were the usual weighted estimates over the two year period. Each diagnosis is represented by two variables, m_d and s_d . The problem is now reduced to finding an appropriate partition of the diagnoses according to m_d and s_d . Three special groups of diagnoses judged as outliers were removed. These three special groups will form the first three rows of the imputation matrix (the columns are defined by the sex variable). A catch-all category was created in the last row of the imputation matrix for those diagnoses with, say, fewer than ten observations available over the two years of data and not included in the three special groups. The choice for the upper bound of ten observations was made arbitrarily. Cluster analysis can then be used to group the remaining diagnoses not included in the three special groups with at least ten observations available.

Step III: Determination of the number of clusters

The determination of the number of clusters was dictated by operational constraints since the imputation module of the program doing the imputation will accept a maximum number of rows not larger than 40. Since there are already three rows for special diagnoses and one row for diagnoses with fewer than ten observations, the maximum number of other rows that would not affect the program is then 36. A small empirical study calculating the R^2 coefficient for different numbers of clusters indicated that the R^2 coefficient was already above 98% for 36 clusters, suggesting that 36 clusters was acceptable. Note that even with 15 clusters, the R^2 could be made as high as 95%. The definition of the R^2 coefficient is given in section 3.4.

Step IV: FASTCLUS implementation

First, an initial partition of the observations into 36 groups was chosen (equivalent to choosing a set of 36 cluster seeds). Better results were obtained by selecting an initial set of cluster seeds than by letting FASTCLUS find initial cluster seeds. Note that different initial cluster seeds and different orders of the input data set will yield different results

due to the fact that the method produces only locally optimal partitions. To select cluster seeds, diagnoses were divided into nine groups of roughly the same size according to m_d and four groups of roughly the same size according to s_d . This procedure produced 36 homogeneous groups of diagnoses of approximately the same size. The means of the two variables m_d and s_d in each group were taken as initial cluster seeds. Several other variations were tried and the procedure giving the largest R^2 was chosen.

Second, since m_d and s_d were based on very different numbers of observations for different diagnoses, it was judged preferable to perform a weighted cluster analysis, the weights being the number of observations available for each diagnosis. Note that, in this case, FASTCLUS would minimize the weighted within cluster sum of squares instead of an unweighted within-cluster sum of squares.

3.4 Relative Performance of the Proposed Method

One way to compare the current and proposed method for collapsing imputation classes is to use the R^2 coefficient pooled over all variables (in our case, it would be the mean and the standard deviation). The pooled R^2 coefficient is the proportion of the total variance explained by the between cluster pooled sum of squares (which should be as large as possible). Each pooled sum of squares is defined as $(SSQ_m + SSQ_s)/2$ where SSQ_m and SSQ_s are the sums of squares of the mean and the standard deviation respectively. The R^2 coefficients obtained from FASTCLUS were 0.993 for m_d and 0.929 for s_d for a pooled R^2 value of 0.986. The current classification of diagnoses into groups would yield an R^2 of 0.735 for m_d and 0.466 for s_d producing a pooled R^2 value of 0.705. Thus, in terms of R^2 , results indicated that the groups of diagnoses formed using cluster analysis were much more homogeneous with respect to the variable being imputed than in the case where classes were formed by the earlier method.

4. CONCLUDING REMARKS

A methodology based on cluster analysis for collapsing the imputation classes of an imputation matrix defined by the cross-classification of several auxiliary variables was proposed. This methodology was applied to the imputation of age for the Hospital Morbidity System where diagnosis and sex were used as auxiliary variables.

It should be noted that in this specific application, only one variable, namely the diagnosis, was used to collapse the original imputation classes. The variable sex is, however, used later in the imputation scheme so that a recipient will be matched to a donor of the same sex. In a generalization of the proposed method, one may consider using the two variables, sex and diagnosis, in the collapsing process. For this purpose one might also impose some constraints that male and female cases of the same diagnosis belong to the same row in the final imputation matrix. Alternatively, one could produce two final imputation matrices, one for each sex. In either one of these alternatives, the number of initial imputation classes would clearly be much higher and hence the collapsing problem more complex. In this situation, it is more likely for many classes to have a small number of donors and therefore many of the imputation classes would have to be assigned to the catch all category. This, however, may not be desirable in practice. This problem can be simplified if one could make the assumption that, for most diagnoses, the male and female age distributions are similar to each other. There is some evidence based on significance tests that this is not an unreasonable assumption. In the HMS example considered, it was decided to group diagnoses based on estimates of μ_d and σ_d from the data pooled over sex.

It should also be noted that the choice of mean and standard deviation of age distribution to assign numerical scores to each imputation class was not investigated. Other choices might be percentiles or some other parameters of the age distribution. Clearly, the results of using cluster analysis for collapsing purpose would depend on the choice of the above scores.

Finally, generalization of the proposed method to the case where $k \geq 1$ variables need to be imputed and where $p \geq 2$ auxiliary variables are available follows in a straightforward manner from the simpler case considered in this paper.

ACKNOWLEDGEMENTS

This work was presented at the annual meeting of the "Association canadienne-française pour l'avancement des sciences" in May 1988. I thank Avi Singh for his helpful comments which have led to improvements in this paper. I would like to thank Cyril Nair of the Health Division and his staff for their support especially concerning the production of the computer files required to complete this work.

REFERENCES

- ANDERBERG, M.R. (1973). *Cluster Analysis for Application*. New York: Academic Press.
- BALL, G.H., and HALL, D.J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics*, 12, 17-31.
- CORMACK, R.M. (1971) A review of classification. *Journal of the Royal Statistical Society, Series A*, 134, 321-367.
- DREW, J.D., BÉLANGER, Y., and FOY, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.
- EVERITT, B.S. (1980). *Cluster Analysis*. Second Edition, London: Heineman Education Books Ltd.
- JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-284.
- HARTIGAN J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.
- MacQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings 5th Berkeley Symposium* 1, 281-297.
- SAS INSTITUTE Inc. (1985). *SAS User's Guide: Statistics*, Version 5.
- STATISTICS CANADA (1986). *Hospital Morbidity 1981-82, 1982-83*. Catalogue No. 82-206, Statistics Canada, Ottawa.
- WORLD HEALTH ORGANIZATION (1977). *International Classification of Diseases*. 1975 Revision, Volume 1, Geneva.

An Example of the Use of Randomization Tests in Testing the Census Questionnaire

YVES BÉLAND and ALAIN THÉBERGE

ABSTRACT

Modular Test 2 was a survey conducted by Statistics Canada that used two different questionnaires. Its purpose was to assist in the making of the 1991 census questionnaire. The sample used for the survey was not a probability sample. This article briefly describes the survey methodology, and the use of randomization tests to compare the two questionnaires.

KEY WORDS: Randomization tests; Non-probability sample; Experimental design.

1. INTRODUCTION

Statistical tests could be classified into two groups, randomization tests and classical tests. A classical test, is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for the set of samples that could have been selected. To conduct this kind of test, the probability of selecting any given sample must be known; therefore probability sampling using a known design is required. A randomization test is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for all possible permutations of the data. This was the method used by Fisher to compare two seed samples (1935), and Edgington (1987) also discusses various aspects of this method. "Treatments" are required to define the permutations in a randomization test, and the probability of obtaining a given permutation must also be known. Which unit will be given which treatment must be decided randomly; that is, the experimental design must incorporate randomization.

In an organization like Statistics Canada, classical tests are generally used because most of the sample surveys done by Statistics Canada use probability sampling, and also because there are no treatments in these surveys. This article describes how randomization tests were used in a survey that was an exception to the rule.

In Section 2, the methodology used in the modular tests is described briefly. Section 3 describes using simple examples the procedure used in a randomization test. Section 4 describes how randomization tests were applied to Modular Test 2.

2. MODULAR TESTS

As part of the planning for the 1991 census, two modular tests were carried out to test questions likely to be asked in the census. The purpose of these surveys was to ensure that each question whether new or just reformulated was easy to understand. We refer to the tests as "modular" because they were independent surveys that tested different sections of the census questionnaire.

¹ Yves Béland, Social Survey Methods Division; Alain Théberge, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

Modular Test 1 was carried out in November, 1987 in order to revise newly-formulated questions dealing with population coverage, marital status, fertility, volunteer work, and nuptiality. This first survey used neither classical nor randomization tests.

Modular Test 2, carried out in January, 1988, was designed principally to measure the reaction of ethnic groups to questions on language, ethnic origin, religion, citizenship, and mobility. In Modular Test 2, a two-stage sampling plan was used to select about 3,500 households taken from within the metropolitan areas of Halifax, Québec, Montréal, Toronto, Winnipeg, and Vancouver. To reduce costs and to make data collection easier, and to get a sample that contained people of diverse ethnic origins, a non-probability method was used to select the sample. The questionnaire used in Modular Test 2 came in two versions. The differences are described in Section 4. The households in the sample were given either version 1 or version 2 on a random basis.

Randomization tests were used to allow us to statistically test hypotheses pertaining to Modular Test 2. Randomizations tests can be used to compare two treatments applied to units in samples which may not be probability samples.

3. RANDOMIZATION TESTS

The procedure for doing a randomization test will now be described. First, the value of a statistic is calculated for the observed data. Next, the value of the same statistic is calculated for the other permutations of the data that are possible with the experimental design used. H_0 is rejected if the value of the statistic for the observed data is extreme in relation to the values obtained under H_0 for the set of permutations.

For example, suppose there are four households. Household 1 has three persons, households 2 and 3 have two, and household 4 has one. These households may have been chosen arbitrarily, but a household whose members will receive treatment Y is chosen at random. Members of the three other households will receive treatment X . Suppose that household 4 is selected for treatment Y . For household 1, the treatment succeeds for two of the three members, for households 2 and 3, for one of two members, and for household 4, it fails for the sole member. Our null hypothesis states that the results are independent of the treatment used. To measure the impact of treatment X compared to treatment Y , the statistic S , giving the average number of successes for treatment X minus the average number of successes for treatment Y is calculated. Here $S = (2 + 1 + 1)/(3 + 2 + 2) - 0/1 = 4/7$. To find out whether this value is significant, the values for S obtained by permuting the observations are given in Table 1. Each observation in Table 1 shows the number of members in the household after the vertical bar, and the number of successes before the vertical bar. If a right-tailed test is used, H_0 is rejected when $\alpha \geq 3/12 = .25$, because three of the twelve permutations yield an S value greater than or equal to $4/7$, the observed value.

Rather than permuting the observations, we could have permuted the treatments. Table 2 gives the results when this is done. Because only one of the four permutations yields a value for S greater than or equal to $4/7$ for a right-tailed test, we again reject H_0 if $\alpha \geq 1/4 = .25$. It is not a coincidence if the results are the same. Note n_{ki} , the number of units that receive treatment k ($k = 1, \dots, K$) and for which the result r_i ($i = 1, \dots, I$) is observed; $n_{k.} = \sum_i n_{ki}$ the number of units that receive treatment k , $n_{.i} = \sum_k n_{ki}$, the number of units for which the result r_i is observed; and $n_{..} = \sum_k \sum_i n_{ki}$, the total number of units. The number, N_t , of permutations of the treatments is given by

$$N_t = n_{..}! / \prod_k (n_{k.}!). \quad (1)$$

Table 1
Values of the Statistics *S* for each Permutation of the Observations

Treatment	Permutations											
<i>X</i>	2 3	1 2	1 2	2 3	2 3	1 2	1 2	0 1	0 1	1 2	1 2	0 1
<i>X</i>	1 2	2 3	1 2	1 2	0 1	2 3	0 1	1 2	2 3	1 2	0 1	1 2
<i>X</i>	1 2	1 2	2 3	0 1	1 2	0 1	2 3	2 3	1 2	0 1	1 2	1 2
<i>Y</i>	0 1	0 1	0 1	1 2	1 2	1 2	1 2	1 2	1 2	2 3	2 3	2 3
<i>S</i>	4/7	4/7	4/7	0	0	0	0	0	0	-4/15	-4/15	-4/15

Table 2
Values of the Statistics *S* for each Permutation of the Treatments

Observation	Permutations			
2 3	<i>X</i>	<i>X</i>	<i>X</i>	<i>Y</i>
1 2	<i>X</i>	<i>X</i>	<i>Y</i>	<i>X</i>
1 2	<i>X</i>	<i>Y</i>	<i>X</i>	<i>X</i>
0 1	<i>Y</i>	<i>X</i>	<i>X</i>	<i>X</i>
<i>S</i>	4/7	0	0	-4/15

Of these N_t permutations, there are N_t^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \dots, K; i = 1, 2, \dots, I$), where

$$N_t^* = \prod_i \left(n_{.i}! / \prod_k (n_{ki}!) \right). \tag{2}$$

In addition, there are N_o permutations of the observations where

$$N_o = n_{..}! / \prod_i (n_{.i}!). \tag{3}$$

Of these N_o permutations, there are N_o^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \dots, K; i = 1, 2, \dots, I$), where

$$N_o^* = \prod_k \left(n_{k.}! / \prod_i (n_{ki}!) \right). \tag{4}$$

Because $N_o^*/N_o = N_t^*/N_t$, the tests are equivalent. To reduce the number of calculations, it is preferable to permute the treatments if $N_t < N_o$, and to permute the observations if $N_t > N_o$. Dwass (1957) suggests that when there are a large number of permutations, a sample of permutations can be taken, and the observed value of the statistic can be compared to the set of values for the sample. If all of the permutations are not considered, the level of the test is not affected, only its power is.

If the permutations are sampled, the rule given above can still be applied, not to reduce the number of calculations, but to minimize the loss of power due to sampling. For example, Dwass shows that for a one-tailed test at the 0.05 level, the loss of power for a sample of 999 permutations is no more than 5.5%. Bradley (1968) notes that when the power of randomization and classical tests are compared, the results depend on to what extent the requirements of the classical tests have been met.

Because of the way in which randomization tests are constructed, the inference applies only to the effect of treatment on units in the sample, and not to the entire population. Classical tests, however, are based on a random sample drawn from a population that rarely matches the population of interest. In the present case for example, the population of interest is the Canadian population on Census Day, June 4, 1991. So for both types of tests, non-statistical arguments must be used to generalize inferences to the population of interest.

4. THE USE OF RANDOMIZATION TESTS IN MODULAR TEST 2

As mentioned above, there are two questionnaire versions for Modular Test 2, versions *X* and *Y*. Questions on ethnic identity and ethnic origin differ in the two versions. "CANADIAN" is a response category in version *X* that the respondent can select to answer the questions on ethnic identity and origin. In version *Y*, those who want to respond "CANADIAN" must write it out in full after selecting the category, "OTHER."

We wanted to know whether questions on ethnic identity and origin in version *X* of the test questionnaire got more or got less multiple responses than these questions in version *Y*. By a multiple response we mean any response in which more than one category has been chosen. We also wanted to find out what bearing the type of questionnaire had on multiplicity (number of response categories selected by the respondent), and on the selection of certain response categories (such as "FRENCH") for these questions. The types of questionnaire constitute the treatments. Because the sample for each region had its peculiarities, the randomization tests were done separately for each of the metropolitan areas from which the sample was taken.

First of all, we generated at random a sample of 999 permutations of the questionnaire versions. A permutation is generated as follows: For any given region, let N_x and N_y represent the number of *X* and *Y* questionnaires respectively. Using Bebbington's algorithm (1975), from the $N_x + N_y$ households take a simple random sample of N_x households. Household members in this sample are then assigned version *X* of the questionnaire. This process is repeated 999 times. Next, calculate for a given question the proportion of respondents who gave a multiple response for version *X* and for version *Y*. These proportions are denoted P_x and P_y .

Next, for each of the 999 permutations of the questionnaire versions, as well as for the initial observed sample, we calculated the statistic $S = P_x - P_y$. In this way we obtained 1,000 values for S , which we ranked in increasing order. If more than one statistic had the same value, we generated a random number between 0 and 1 and used it to determine the order of statistics of the same value. We used the variable $RANKP_{x-y}$ to represent the rank of an observed S statistic.

Let μ_x and μ_y represent the expected proportion of respondents who gave a multiple response for version *X* and version *Y* respectively. For all regions excluding Halifax we tested:

$$H_0: \mu_x = \mu_y$$

versus

$$H_1: \mu_x > \mu_y.$$

For Halifax, the counter-hypothesis $H_1: \mu_x < \mu_y$ was used because more multiple responses were expected for version Y of the questionnaire. Because "CANADIAN" was not an available response category on version Y of the questionnaire and because the majority of households selected in this region were made up of people of British origin (that is, English, Scottish, or Irish), members of households that received version Y marked one or more of these categories. Members of households that received version X had the option of marking only the "CANADIAN" category.

The critical level, $\hat{\alpha}$, is calculated as follows: for the Halifax region, given that H_0 is rejected if the proportion of respondents who gave a multiple response in version X is significantly lower than the proportion observed for Y , the critical level is $\text{RANKP}_{x-y} / 1000$. For all the other regions, given that H_0 is rejected if the proportion for X is significantly higher than the proportion observed for Y , the critical level is $(1001 - \text{RANKP}_{x-y}) / 1000$. The results are shown in Table 3.

Randomization tests were also used to test multiplicity (the number of response categories selected by the respondent) for questions on ethnic identity and origin in each of the regions, but this time ratios (R_x, R_y) are used, instead of proportions (P_x, P_y). Ratio R_x is the average number of response categories selected by respondent for a question in version X of the questionnaire, and ratio R_y is the average number of response categories selected in version Y . The rest of the method is the same except that instead of RANKP_{x-y} , RANKR_{x-y} is used, and the statistic S is defined as $R_x - R_y$. However, because there is greater variability for the values of the statistic S in the tests for multiplicity, a sample of 1,999 permutations was generated instead of 999.

Let F and G represent the distribution functions of the number of response categories selected in version X and version Y respectively. For all the regions excluding Halifax, we test the hypothesis

$$H_0: F = G$$

versus

$$H_1: F(z) \leq G(z) \text{ for all } z \text{ and } F \neq G.$$

If H_0 is rejected, the number of response categories selected for an X questionnaire is said to be stochastically larger than the number of response categories selected for a Y questionnaire. For Halifax, the counter-hypothesis used is $H_1: F(z) \geq G(z)$, for all z and $F \neq G$. The results are shown in Table 3. In the Québec region, the value of R_y is less than 1 for each question. This is because most respondents in this region chose only one response category, and some respondents did not answer one or other of the questions.

Finally, versions X and Y for Modular Test 2 were compared for some regions as to the number of respondents who identified themselves as being of French, Italian, or British origin. By "BRITISH", we mean that at least one of the categories "IRISH," "SCOTTISH," or "ENGLISH" was chosen. For example, if a test was done on the proportion of people selecting "FRENCH", μ_x and μ_y were defined as the expected proportion of questionnaires where the response "FRENCH" would be chosen in versions X and Y of the questionnaire. In all regions, we tested

$$H_0: \mu_x = \mu_y$$

versus

$$H_1: \mu_x < \mu_y.$$

The randomization tests were done using 999 permutations. The results are shown in Table 4.

Table 3
Critical Levels for the Rate of Multiple Responses and Multiplicity

Question	Region	Multiple Response			Multiplicity		
		P_x	P_y	$\hat{\alpha}$	R_x	R_y	$\hat{\alpha}$
ORIGIN	HALIFAX	0.435	0.536	0.087	1.617	1.914	0.062
ORIGIN	QUÉBEC	0.154	0.043	0.001	1.143	0.986	0.001
ORIGIN	MONTRÉAL	0.185	0.194	0.612	1.141	1.152	0.585
ORIGIN	TORONTO	0.127	0.122	0.393	1.124	1.125	0.495
ORIGIN	WINNIPEG	0.293	0.307	0.622	1.439	1.398	0.345
ORIGIN	VANCOUVER	0.285	0.296	0.621	1.440	1.392	0.280
IDENTITY	HALIFAX	0.220	0.335	0.035	1.244	1.502	0.029
IDENTITY	QUÉBEC	0.140	0.016	0.001	1.131	0.959	0.001
IDENTITY	MONTRÉAL	0.159	0.125	0.063	1.075	1.044	0.186
IDENTITY	TORONTO	0.186	0.120	0.001	1.154	1.075	0.005
IDENTITY	WINNIPEG	0.224	0.195	0.248	1.253	1.208	0.298
IDENTITY	VANCOUVER	0.186	0.183	0.457	1.182	1.137	0.202

Table 4
Critical Levels for Selected Variables

Question	Variable	Region	P_x	P_y	$\hat{\alpha}$
ORIGIN	FRENCH	QUÉBEC	0.127	0.897	0.001
ORIGIN	FRENCH	MONTRÉAL	0.038	0.210	0.001
ORIGIN	BRITISH	HALIFAX	0.321	0.837	0.001
ORIGIN	BRITISH	MONTRÉAL	0.034	0.092	0.002
ORIGIN	BRITISH	TORONTO	0.085	0.135	0.003
ORIGIN	BRITISH	WINNIPEG	0.167	0.234	0.054
ORIGIN	BRITISH	VANCOUVER	0.267	0.325	0.065
IDENTITY	FRENCH	QUÉBEC	0.138	0.899	0.001
IDENTITY	BRITISH	HALIFAX	0.153	0.828	0.001
IDENTITY	BRITISH	MONTRÉAL	0.022	0.117	0.001
IDENTITY	BRITISH	TORONTO	0.050	0.215	0.001
IDENTITY	BRITISH	WINNIPEG	0.074	0.276	0.001
IDENTITY	BRITISH	VANCOUVER	0.104	0.325	0.001
IDENTITY	ITALIAN	TORONTO	0.412	0.463	0.060

5. CONCLUSION

The results for tests on the rate of multiple responses are similar to those on multiplicity, which is not surprising. When you compare the critical levels for the question on ethnic origin to the critical levels for the question on ethnic identity, it is seen that the differences between the two versions of the questionnaire affect the responses to the question on ethnic identity the most.

Our main reason for using randomization tests was that the sample for Modular Test 2 was a non-probability sample. However, there are also other cases where randomization tests are appropriate. For example, to do a "Student's" t test for means equality the hypothesis of normality is required, and it must also be assumed that the variances are equal. These assumptions are not needed for a randomization test. It should be kept in mind that the results of a randomization test apply to the sample, and not necessarily to the entire population, unless a simple random sample is used.

REFERENCES

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- BRADLEY, J.V. (1968). *Distribution-free Statistical Tests*. Englewood Cliffs: Prentice-Hall.
- DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.
- EDGINGTON, E.S. (1987). *Randomization Tests*, (2nd ed.). New York: Marcel Dekker.
- FISHER, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Variance Formulae for Composite Estimators in Rotation Designs

PATRICK J. CANTWELL¹

ABSTRACT

In many government surveys, respondents are interviewed a set number of times during the life of the survey, a practice referred to as a rotation design or repeated sampling. Often composite estimation – where data from the current and earlier periods of time are combined – is used to measure the level of a characteristic of interest. As other authors have observed, composite estimation can be used in a rotation design to decrease the variance of estimators of change in level. In this paper, simple expressions are derived for the variance of a general class of composite estimators for level, change in level, and average level over time. Considered first are “one-level” rotation designs, where only the current month is referenced in the interview. Results are developed for any sampling pattern of m interviews over a period of M months. Subsequently, “multi-level” plans are addressed. In each month one of p different groups is interviewed. Respondents then answer questions referring to the previous p months. Results from the several sections apply to a wide range of government surveys.

KEY WORDS: Repeated sampling in surveys; Balanced designs; Month-to-month change; Yearly average.

1. INTRODUCTION

Rotation designs of various types are used in many major household surveys. The Current Population Survey (CPS) is conducted by the U.S. Bureau of the Census for the U.S. Bureau of Labor Statistics. Statistics Canada operates the Labour Force Survey (LFS). Both surveys yield estimates of labor force characteristics, including unemployment. In each survey, households are interviewed a number of times before leaving the sample. In the CPS, each household is “rotated in” for interviews in four consecutive months, rotated out of the sample for eight months, and finally back in for four more months. In the LFS, a participating household responds for six consecutive months and does not return.

A survey with a rotation design lies somewhere between a fixed panel survey, where participants remain in sample indefinitely, and a survey using independent samples, where respondents are interviewed once and retired from sample. The total overlap of a fixed panel from one time period to the next can minimize the variance of estimators of change when measurements are positively correlated across periods. Also, certain costs are incurred only the first time a unit is placed in sample. However, response burden on the members of a fixed panel can be excessive. Using a rotation design is an attempt to realize variance or cost reductions without overly burdening sample participants. In the CPS and the LFS, there are sample overlaps of 75% and 83%, respectively, from one month to the next. For more on these topics, see Woodruff (1963), Rao and Graham (1964), or Wolter (1979).

Some estimators used with rotation designs are composite in nature. In order to take advantage of repeated sampling, they combine rotation group estimates obtained for the current month with those from prior months into a final estimator.

¹ Patrick J. Cantwell, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

While the variance of composite estimators can be decreased by selecting the combination wisely, calculating this variance may become more complex because of the correlation patterns involved among the repeated groups. For general rotation plans, subject to specific restrictions, simple formulae are presented in this paper for the variance of estimators of level and change. The derivations are applied to an important and quite general class of estimators called the generalized composite estimator (Breau and Ernst 1983).

These formulae can be of use if the correlations between estimates from the same rotation group one or more time periods apart can be estimated and are sufficiently large to render composite estimation worthwhile. In continuing government surveys, past sample data will typically enable the estimation of these correlations. Characteristics involving household income and labor force usually exhibit moderately high correlations. For others, such as the incidence of crime, however, correlations across time periods may not be large enough to realize the benefits of composite estimation. Of the surveys mentioned in this paper, only CPS currently uses a composite estimator.

In the developments which follow, two types of surveys are treated separately. In surveys such as CPS and LFS, participants supply information only for the current month. Such surveys are called "one-level" surveys. On the other hand, the U.S. Census Bureau conducts the Survey of Income and Program Participation (SIPP) to acquire data on income level, sources of income, program participation, and other items. During each interview, respondents in the SIPP refer back to the previous four months. A different group is then interviewed the following month. The SIPP design is consequently called "multi-level." The level of a survey was used by Wolter (1979) to indicate the number of periods for which information is solicited in one interview.

Another distinction is made between these two types of surveys. Let the term "design gap" indicate a period of time between interviews which is never referenced in any interview. While the LFS contains no design gaps, CPS includes one of eight months. For the sections pertaining to one-level designs, the results and derivations apply *regardless of the pattern of interviews and design gaps*. Therefore, the formulae are relevant not only to the current design of CPS and LFS, but also to other designs under consideration.

For reasons discussed later, design gaps are generally not a feature of multi-level rotation plans in practice. The SIPP is no exception. Accordingly, the multi-level plans addressed in this paper do not include design gaps.

One-level designs are treated in Sections 2 and 3. In Section 2, the generalized composite estimator is defined. Notation, definitions and covariance assumptions are introduced. The main results – Theorems 1 through 3 – are given in Section 3. Variances of estimators of level and change in level are stated. The formulae are determined for single time periods (such as months) and combinations (such as quarters or years). They apply to one-level designs with any pattern of interviews and design gaps. When seeking the optimal rotation plan and composite estimator, the user must determine how best to combine variance reductions/increases for the resulting estimators of level, "month-to-month" change, and average over many periods.

In Section 4, these results are extended from one-level to certain multi-level designs, which include the SIPP. Subject to minor restrictions – in particular, the exclusion of design gaps in the sampling scheme – theorems similar to those in Section 3 are stated. Because the derivations are analogous to those for one-level plans, the results are not proved.

2. ONE-LEVEL DESIGNS: NOTATION AND DEFINITIONS

Although rotation schemes can assume infinitely many forms, the discussion in Sections 2 and 3 is restricted to one type. At each period of time, a new rotation group enters the sample,

and follows the same pattern of interviews and design gaps as every preceding group. In addition, responses refer only to the current period of time, whether or not the participants were in sample in the previous period. This design is called a balanced one-level rotation plan. The design is "balanced" because the number of groups in sample at any time is equal to the total number of time periods any one group is included in the sample.

The scheme used in the LFS satisfies these restrictions. Each month a new group enters, and remains in the sample for five more months. The CPS as it currently operates follows these guidelines in a 4-8-4 plan. Before July 1953, however, CPS used an unbalanced design where five rotation groups entered, one each in consecutive months. In the sixth month, *no new group entered*. The process then continued in the same manner, with groups exiting after six months in sample.

One problem with the CPS design before 1953 is the introduction of month-in-sample bias, often referred to as rotation group bias. Of greater concern here is the changing pattern of rotation group appearances. The variance of a composite estimate depends on when each participating group appeared in sample before, and the covariance structure for identical groups in different months. If the pattern of appearances changes from month to month, the variance formula of the estimator also changes. Under a balanced design with stationary covariance structure, general derivations are possible.

Throughout this paper, the word "month" refers to the period of time in which interviews are done, partly for brevity, but also because most government surveys use the month to divide the life of the survey. However, the results in this section and the next apply to any period of time, provided the rotation plan is balanced and one-level.

Some notation and vector definitions are now introduced. Suppose that every rotation group is in sample for a total of m interviews over a period of M months. That is, it is out of sample for $M - m$ months after first entering and before exiting. The balanced design ensures that m groups are in sample during any month.

The set T_0 is defined as follows. Consider any rotation group. Let T_0 index the set of "months" when this group is *not in sample*, labeling as month one the month this group is first interviewed, and stopping at month M . Because the design is balanced, the composition of T_0 does not depend on which group is selected. Note that, if respondents are interviewed in m consecutive months, *i.e.*, there are no design gaps, then m and M are the same, and T_0 is empty.

Next, given a set of m values w_1, \dots, w_m , it is possible to define the $M \times 1$ vector w as follows. Define the i th component of w to be 0 if $i \in T_0$. This step fills $M - m$ positions in w . Then the values w_1, \dots, w_m are inserted in order into the remaining m components, starting with the first. The resulting w is called a vector "in design form." For example, in a 4-8-4 rotation plan, $T_0 = \{5, 6, \dots, 12\}$, and $w^T = (w_1, w_2, w_3, w_4, 0, 0, 0, 0, 0, 0, 0, 0, w_5, w_6, w_7, w_8)$.

It is useful to introduce the $M \times M$ matrix R as: $R_{ii} = 1$ if $i \notin T_0$, and 0 if $i \in T_0$; and $R_{ij} = 0$ if $i \neq j$. It is clear that R is a diagonal matrix where $\text{diag}(R)$ is a set of 1's "in design form," R_{11} and R_{MM} are 1, and $\sum_{i=1}^M R_{ii} = m$.

Observe that, for any $M \times p$ matrix V , RV is the same as V , but with 0's across each row i such that i is in T_0 . In other words, premultiplication by R "removes" (turns to 0) the rows of V indexed by T_0 . If the columns of V are already in design form, then $RV = V$. Similarly, for any $p \times M$ matrix U , postmultiplication by R "removes" the columns of U which are indexed by T_0 . If the rows of U are already in design form, then $UR = U$.

Let L be the $M \times M$ matrix with 1's on the subdiagonal, and 0's elsewhere. Formally, $L_{ij} = 1$, if $i - j = 1$, and 0, otherwise. For any $M \times 1$ vector written as $w^T = (w_1, \dots, w_M)$, the product Lw becomes $(0, w_1, w_2, \dots, w_{M-1})^T$, and $w^T L$ is $(w_2, w_3, \dots, w_M, 0)$.

Turning to the data, let $x_{h,i}$ denote the estimate of "monthly" level for some characteristic to be measured from the rotation group which is in sample for the i th time in month h , where $i = 1, \dots, m$. Breau and Ernst (1983) defined the generalized composite estimator (GCE) of level recursively as follows. For monthly level, let:

$$y_h = \sum_{i=1}^m a_i x_{h,i} - k \sum_{i=1}^m b_i x_{h-1,i} + k y_{h-1}, \quad (1)$$

where $0 \leq k < 1$, and the a_i 's and b_i 's may take any values, including negative ones, subject to $\sum_{i=1}^m a_i = 1$ and $\sum_{i=1}^m b_i = 1$. The "current composite" and AK composite estimators used in CPS are special cases of the GCE. For information on these, see Hanson (1978), Huang and Ernst (1981), and Kumar and Lee (1983).

The GCE is more restrictive than a general linear estimator which combines $x_{h,i}$ values from the current period with those from many prior months (see Gurney and Daly 1965). However, the GCE has been shown to perform almost as well (Breau and Ernst 1983). It has the advantage that only data from two months – the current month and the preceding one – need to be stored. Although y_h incorporates earlier data, it is summarized through y_{h-1} .

To facilitate variance computations, (1) is expressed in vector form. Let \mathbf{a} and \mathbf{b} be $M \times 1$ vectors in design form comprising, respectively, the sets of constants a_1, \dots, a_m and b_1, \dots, b_m . Similarly, for any h , the observations $x_{h,1}, \dots, x_{h,m}$ make up \mathbf{x}_h , also an $M \times 1$ vector in design form. Then

$$y_h = \mathbf{a}^T \mathbf{x}_h - k \mathbf{b}^T \mathbf{x}_{h-1} + k y_{h-1}. \quad (1a)$$

The data are assumed to exhibit a stationary covariance structure:

- (i) $\text{Var}(x_{h,i}) = \sigma^2$ for all h and i ;
- (ii) $\text{Cov}(x_{h,i}, x_{h,j}) = 0$ for $i \neq j$, i.e., different rotation groups in the same month are uncorrelated; and
- (iii) $\text{Cov}(x_{h,i}, x_{s,j}) = \rho_{|h-s|} \sigma^2$, if the two x 's refer to the same rotation group $|h-s|$ months apart; or 0, otherwise. Take ρ_0 to be 1. (2)

From the first two parts of (2), it is clear that $\text{Var}(\mathbf{x}_h) = \sigma^2 \mathbf{R}$, for all h . Part three implies that $\text{Cov}(\mathbf{x}_h, \mathbf{x}_{h-1}) = \sigma^2 \rho_1 \mathbf{R} \mathbf{L} \mathbf{R}$. This follows because (a) the matrix \mathbf{L} , with 1's on the sub-diagonal, "represents" the one month lag between the \mathbf{x}_h and \mathbf{x}_{h-1} values, and (b) pre-multiplying (postmultiplying) by \mathbf{R} inserts 0's corresponding to 0's in \mathbf{x}_h (\mathbf{x}_{h-1}) (months not in sample).

It is readily seen that $(\mathbf{L}^r)_{ij} = 1$ if $i - j = r \geq 0$ and $1 \leq j, i \leq M$; take \mathbf{L}^0 to be the identity matrix. The same development as above gives $\text{Cov}(\mathbf{x}_h, \mathbf{x}_{h-2}) = \sigma^2 \rho_2 \mathbf{R} \mathbf{L}^2 \mathbf{R}$. In general,

$$\text{Cov}(\mathbf{x}_h, \mathbf{x}_{h-r}) = \sigma^2 \rho_r \mathbf{R} \mathbf{L}^r \mathbf{R}, \text{ for } r = 0, 1, 2, \dots, \text{ and all } h. \quad (3)$$

For $r \geq M$, $\mathbf{L}^r = 0$, and $\text{Cov}(\mathbf{x}_h, \mathbf{x}_{h-r}) = 0$.

For the theorems which follow, define the $M \times M$ matrix \mathbf{Q} by: $Q_{ij} = k^{i-j} \rho_{i-j}$, if $1 \leq j < i \leq M$, and 0, otherwise. Finally, let \mathbf{I} be the $M \times M$ identity matrix.

3. ONE-LEVEL DESIGNS: THEOREMS AND PROOFS

Three theorems are now stated and proved.

Theorem 1. If the GCE of level is defined as in (1), and the covariance structure as expressed in (2) holds, then

$$\text{Var}(y_h) = \sigma^2 \{ a^T a + k^2 b^T (b - 2a) + 2(a - k^2 b)^T Q (a - b) \} / (1 - k^2). \quad (4)$$

Notice that when one uses an unweighted average of the estimates from the m rotation groups of the current month, $k = 0$, $Q = 0$, and $a_i = 1/m$, for $i = 1, \dots, m$. Then $\text{Var}(y_h) = \sigma^2/m$, as expected.

Proof of theorem 1. Substitution into (1a) recursively leads to

$$y_h = a^T x_h + (a - b)^T \sum_{i=1}^{\infty} k^i x_{h-i}. \quad (5)$$

From (3), the variance of this sum is

$$\begin{aligned} \text{Var}(y_h) &= a^T \sigma^2 R a + (a - b)^T \sum_{i=1}^{\infty} k^{2i} \sigma^2 R (a - b) \\ &\quad + 2a^T \sum_{i=1}^{\infty} k^i \sigma^2 \rho_i R L^i R (a - b) \\ &\quad + 2(a - b)^T \sum_{i=1}^{\infty} \sum_{j=i+1}^{\infty} k^{i+j} \sigma^2 \rho_{j-i} R L^{j-i} R (a - b) \\ &= \sigma^2 \left\{ a^T R a + (a - b)^T R (a - b) k^2 / (1 - k^2) \right. \\ &\quad + 2a^T R \left(\sum_{i=1}^{\infty} k^i \rho_i L^i \right) R (a - b) \\ &\quad \left. + 2(a - b)^T R \left(\sum_{i=1}^{\infty} k^{2i} \left[\sum_{j=i+1}^{\infty} k^{j-i} \rho_{j-i} L^{j-i} \right] \right) R (a - b) \right\}. \quad (6) \end{aligned}$$

Because a and $a - b$ are vectors in design form, $a^T R = a^T$, $(a - b)^T R = (a - b)^T$, and $R(a - b) = (a - b)$. The sum $\sum_{i=1}^{\infty} k^i \rho_i L^i$ is seen to be the matrix Q : its ij th entry is $k^{i-j} \rho_{i-j}$, if $1 \leq j < i \leq M$, and 0, otherwise. A change of variables will show that the sum in brackets is also Q . Expression (6) can be rewritten as:

$$\begin{aligned} &\sigma^2 \{ a^T a + (a - b)^T (a - b) k^2 / (1 - k^2) + 2a^T Q (a - b) \\ &\quad + 2(a - b)^T Q (a - b) k^2 / (1 - k^2) \}. \end{aligned}$$

Simple rearrangement of these terms produces the result in (4).

Theorem 2. Let $y_h - y_{h-1}$ be the GCE estimator of “month-to-month” change. Then $\text{Var}(y_h - y_{h-1})$ is

- (i) $2\sigma^2 \mathbf{a}^T (\mathbf{I} - \rho_1 \mathbf{L}) \mathbf{a}$, if $k = 0$, and
- (ii) $\sigma^2 (\mathbf{a}^T \mathbf{a} + k^2 \mathbf{b}^T \mathbf{b} - 2k\rho_1 \mathbf{a}^T \mathbf{L} \mathbf{b}) / k - (1 - k)^2 \text{Var}(y_h) / k$, if $0 < k < 1$.

Proof of theorem 2:

- (i) If $k = 0$, $y_h = \mathbf{a}^T \mathbf{x}_h$. From (3), the variance of $\mathbf{a}^T \mathbf{x}_h - \mathbf{a}^T \mathbf{x}_{h-1}$ is

$$2\mathbf{a}^T \sigma^2 \mathbf{R} \mathbf{a} - 2\mathbf{a}^T \sigma^2 \rho_1 \mathbf{R} \mathbf{L} \mathbf{R} \mathbf{a} = 2\sigma^2 \mathbf{a}^T (\mathbf{I} - \rho_1 \mathbf{L}) \mathbf{a}.$$

- (ii) If $0 < k < 1$, define W_h as $\mathbf{a}^T \mathbf{x}_h - k\mathbf{b}^T \mathbf{x}_{h-1}$. From prior results, it is quickly seen that

$$\text{Var}(W_h) = \sigma^2 \{ \mathbf{a}^T \mathbf{a} + k^2 \mathbf{b}^T \mathbf{b} - 2k\rho_1 \mathbf{a}^T \mathbf{L} \mathbf{b} \}. \quad (7)$$

From (1a), $y_h = W_h + ky_{h-1}$. Then

$$\text{Var}(y_h) = \text{Var}(W_h) + k^2 \text{Var}(y_{h-1}) + 2k \text{Cov}(W_h, y_{h-1}); \quad (8)$$

the covariance term can be isolated for later use. Finally, $y_h - y_{h-1} = W_h - (1 - k)y_{h-1}$. When computing the variance of this difference, substitution from (8) and (7) produces the desired result.

Often of primary importance are the average level over a certain length of time (e.g., a quarter or a year), the difference in these averages from one “year” to the next, or the difference in “monthly” level for two months a year apart. Denote by $S_{h,t}$ the sum of the GCE’s for the last t months:

$$S_{h,t} = y_h + y_{h-1} + \dots + y_{h-t+1}, \quad t \geq 1. \quad (9)$$

Commonly used values of t include three, four and twelve. It is left to the reader to divide $S_{h,t}$ by t if an average desired rather than a sum.

Theorem 3:

(a) The expressions $S_{h,t}$, $S_{h,t} - S_{h-t,t}$, and $y_h - y_{h-t}$ can be written as $\sum_{i=0}^{\infty} v_i^T \mathbf{x}_{h-1}$, where

- (i) for $S_{h,t}$, $v_i =$

$$\begin{aligned} &\mathbf{a} + [(k - k^{i+1}) / (1 - k)] (\mathbf{a} - \mathbf{b}), \quad \text{for } i = 0, 1, \dots, t-1, \\ &[k^{i-t} (k - k^{t+1}) / (1 - k)] (\mathbf{a} - \mathbf{b}), \quad \text{for } i = t, t+1, t+2, \dots; \end{aligned}$$

- (ii) for $S_{h,t} - S_{h-t,t}$, $v_i =$

$$\begin{aligned} &\mathbf{a} + [(k - k^{i+1}) / (1 - k)] (\mathbf{a} - \mathbf{b}), \quad \text{for } i = 0, 1, \dots, t-1, \\ &[(2k^{i-t+1} - k - k^{i+1}) / (1 - k)] (\mathbf{a} - \mathbf{b}) - \mathbf{a}, \quad \text{for } i = t, t+1, \dots, 2t-1, \\ &- [k^{i-2t+1} (1 - k^t)^2 / (1 - k)] (\mathbf{a} - \mathbf{b}), \quad \text{for } i = 2t, 2t+1, \dots; \end{aligned}$$

- (iii) for $y_h - y_{h-t}$, $v_0 = \mathbf{a}$, $v_t = k^t (\mathbf{a} - \mathbf{b}) - \mathbf{a}$, and $v_i =$

$$\begin{aligned} &k^i (\mathbf{a} - \mathbf{b}), \quad \text{for } i = 1, 2, \dots, t-1, \\ &- k^{i-t} (1 - k^t) (\mathbf{a} - \mathbf{b}), \quad \text{for } i = t+1, t+2, \dots; \end{aligned} \quad (10)$$

(b) For the sets of vectors v_0, v_1, v_2, \dots defined in (a),

$$\text{Var} \left(\sum_{i=0}^{\infty} v_i^T x_{h-i} \right) = \sigma^2 \left\{ \sum_{i=0}^{\infty} v_i^T v_i + 2 \sum_{i=0}^{\infty} v_i^T \sum_{n=1}^{M-1} \rho_n L^n v_{i+n} \right\}; \quad (11)$$

the sums in (11) converge.

Proof of theorem 3. For (a), successive inclusion of terms y_h through y_{h-t+1} , and the application of (5) to y_{h-t} yield

$$\begin{aligned} S_{h,t} &= a^T(x_h + x_{h-1} + \dots + x_{h-t+1}) + k(a - b)^T x_{h-1} \\ &\quad + (k + k^2)(a - b)^T x_{h-2} + \dots \\ &\quad + (k + k^2 + \dots + k^{t-1})(a - b)^T x_{h-t+1} \\ &\quad + (k + k^2 + \dots + k^t)(a - b)^T \sum_{j=t}^{\infty} k^{j-t} x_{h-j}. \end{aligned} \quad (12)$$

The three sets of v_i 's are then determined from (12) and (5).

The proof of (b) is similar to that of Theorem 1, once it is seen that the v_i 's defined in (a), being linear combinations of a and $a - b$, are in design form. To prove convergence, note that, for all three sets of v_i 's in (a), v_i is proportional to $k^i(a - b)$ for i sufficiently large. There exists a constant $\lambda > 0$ such that, for $i \geq 2t$ and each component j , $|v_{ij}| \leq k^i \lambda$. Recalling that $|\rho_i| \leq 1$, and that each row of L^n has at most one nonzero element (equal to 1), the finite sum in (11) is seen to be an $M \times 1$ vector, each of whose components is bounded above in absolute value by $k^i(M - 1)\lambda$. Convergence of the double summation then follows geometrically in k^{2i} .

4. EXTENSION TO MULTI-LEVEL DESIGNS

Although the results developed in Sections 2 and 3 apply to all balanced one-level rotation plans, it was observed that many surveys operate under multi-level designs. For example, in the Survey of Income and Program Participation (SIPP), one of four rotation groups is interviewed each month, and respondents supply information about the previous four months. Although the design is always subject to change, the first rotation group is interviewed in February, June, October, February, *etc.*, for a total of eight interviews. A second group is interviewed in March, July, *etc.* The remaining two groups follow the same sampling pattern, beginning in April and May. A SIPP panel is the set of four concurrent rotation groups covering about two and one-half years. Each year, a new panel is introduced. For example, the 1986 panel ran from 1986 through 1988, while the 1987 panel spanned 1987-89. Data from different panels are not combined, even though they may cover a common year or two. For further details on the SIPP design, see Nelson, McMillen and Kasprzyk (1984).

When one-level designs were addressed, a rotation group was allowed to assume any pattern of interviews and design gaps – intermediate months which are never referenced – provided the design was balanced. In a multi-level plan, however, design gaps can create problems with recall. Looking back several months, a respondent may find it difficult to assign an event to

the correct period of time. Design gaps can only add to the confusion. For this reason, and because multi-level surveys which incorporate design gaps are rare in practice, this section considers only designs where (i) the sample comprises p rotation groups, (ii) groups are interviewed every p th “month” in an alternating sequence, and (iii) the period of reference is the previous p months.

Many multi-level surveys, for example, the National Crime Survey, sponsored by the U.S. Bureau of Justice Statistics, have a more intricate rotational pattern than that covered here. As expected, variance formulae applied to composite estimators would tend to be more complex.

The interview of a rotation group will refer to the collective gathering of information in the assigned month from all sample units in that group. For a particular characteristic which is to be estimated, let $x_{h,i}$ denote the estimate of “monthly” level for month h from the group which is interviewed in month $h + i$, where $i = 1, \dots, p$. The index i measures recall time – the amount of time between the month of reference and the interview. Table 1 depicts the estimates $x_{h,i}$ for a four-group four-level design. In the diagram solid lines separate estimates which are obtained in different interviews. These boundaries between the reference periods of consecutive interviews are called “seams” in the SIPP.

Table 1
Layout of Estimates in a Longitudinal 4-Level Design

MONTH ↓	ROTATION GROUPS →	1	2	3	4
1		<div>$x_{1,4}$</div>			
2		<div>$x_{2,3}$</div>	<div>$x_{2,4}$</div>		
3		<div>$x_{3,2}$</div>	<div>$x_{3,3}$</div>	<div>$x_{3,4}$</div>	
4		<div>$x_{4,1}$</div>	<div>$x_{4,2}$</div>	<div>$x_{4,3}$</div>	<div>$x_{4,4}$</div>
5		<div>$x_{5,4}$</div>	<div>$x_{5,1}$</div>	<div>$x_{5,2}$</div>	<div>$x_{5,3}$</div>
6		<div>$x_{6,3}$</div>	<div>$x_{6,4}$</div>	<div>$x_{6,1}$</div>	<div>$x_{6,2}$</div>
7		<div>$x_{7,2}$</div>	<div>$x_{7,3}$</div>	<div>$x_{7,4}$</div>	<div>$x_{7,1}$</div>
8		<div>$x_{8,1}$</div>	<div>$x_{8,2}$</div>	<div>$x_{8,3}$</div>	<div>$x_{8,4}$</div>
9		<div>$x_{9,4}$</div>	<div>$x_{9,1}$</div>	<div>$x_{9,2}$</div>	<div>$x_{9,3}$</div>
10		<div>$x_{10,3}$</div>	<div>$x_{10,4}$</div>	<div>$x_{10,1}$</div>	<div>$x_{10,2}$</div>
11		<div>$x_{11,2}$</div>	<div>$x_{11,3}$</div>	<div>$x_{11,4}$</div>	<div>$x_{11,1}$</div>
12		<div>$x_{12,1}$</div>	<div>$x_{12,2}$</div>	<div>$x_{12,3}$</div>	<div>$x_{12,4}$</div>
13		<div>$x_{13,4}$</div>	<div>$x_{13,1}$</div>	<div>$x_{13,2}$</div>	<div>$x_{13,3}$</div>
14		<div>$x_{14,3}$</div>	<div>$x_{14,4}$</div>	<div>$x_{14,1}$</div>	<div>$x_{14,2}$</div>
.	
.	
.	

Note: $x_{h,i}$ denotes the estimate of “monthly” level for month h from the group which is interviewed in month $h + i$. Interviewing begins in month 5. Solid horizontal lines (seams) separate estimates which are obtained in different interviews.

Let the vector \mathbf{x}_h , defined as $(x_{h,1}, x_{h,2}, \dots, x_{h,p})^T$, comprise the p estimates for month h obtained from the p groups in different interviews. Note that $x_{h,p}, x_{h+1,p-1}, \dots, x_{h+p-1,1}$ are estimates for p different months obtained from one group in a single interview (in month $h + p$).

As in Sections 2 and 3, the generalized composite estimator for monthly level is defined as

$$y_h = \sum a_i x_{h,i} - k \sum b_i x_{h-1,i} + k y_{h-1}, \quad (13)$$

where the summations now range from 1 to p . Defining \mathbf{a} and \mathbf{b} as $(a_1, \dots, a_p)^T$ and $(b_1, \dots, b_p)^T$, respectively, the GCE can again be written as

$$y_h = \mathbf{a}^T \mathbf{x}_h - k \mathbf{b}^T \mathbf{x}_{h-1} + k y_{h-1}.$$

The covariance structure of the monthly rotation group estimates is assumed to be stationary in time. Under this multi-level design, however, the length of time between the target month h and the corresponding interview in month $h + i$ may affect the variability of the response, $x_{h,i}$. For $i = 1, \dots, p$, let d_i^2 represent the response variability as a function of the amount of time between the reference month and the interview. The following covariance structure is postulated:

- (i) $\text{Var}(x_{h,i}) = d_i^2 \sigma^2$ for all h and i , where $d_i > 0$;
- (ii) $\text{Cov}(x_{h,i}, x_{h,j}) = 0$ for $i \neq j$; and
- (iii) For $r \geq 0$: $\text{Cov}(x_{h,i}, x_{h-r,j}) = \rho_{r,i} d_i d_j \sigma^2$, if the two x 's refer to the same group r months apart; or 0, otherwise. Take $\rho_{0,i}$ to be 1 for all i . (14)

It may well be that $d_1 \leq d_2 \leq \dots \leq d_p$, if response variability increases with recall time. The subscript r in the correlation coefficient $\rho_{r,i}$ is the amount of time between the months referenced by estimates $x_{h,i}$ and $x_{h-r,j}$. The subscript i indicates that the estimate for month h is obtained from an interview i months later. For specified values of h, r and i , there is only one value j , $1 \leq j \leq p$, for which the estimates $x_{h,i}$ and $x_{h-r,j}$ refer to the same panel and $\text{Cov}(x_{h,i}, x_{h-r,j})$ is nonzero. (This value is $j = \text{mod}_p(i + r - 1) + 1$, where $\text{mod}_p(n)$ is the value of the integer n , modulo p .) Otherwise, the covariance is 0. In some cases, it may be appropriate to replace $\rho_{r,1}, \dots, \rho_{r,p}$ with a common ρ_r .

No assumptions are made about bias. In addition to the effect of recall on variances of group estimates as postulated in (14), a bias related to recall time might also be incurred. Another source - time-in-sample bias - can result according to the number of times a respondent has been interviewed (Bailar 1975). Although these biases need not be measured to derive the variance formulae given in this section, they might constitute a nontrivial component of mean squared error.

Define the $p \times p$ matrices \mathbf{D} , \mathbf{P}_r and \mathbf{J} as follows. Let \mathbf{D} and \mathbf{P}_r , for $r \geq 0$, be diagonal matrices with d_1, \dots, d_p and $\rho_{r,1}, \dots, \rho_{r,p}$, respectively, along the diagonal. Define \mathbf{J} as: $J_{i,i+1} = 1$ for $i = 1, 2, \dots, p - 1$; $J_{p,1} = 1$; and $J_{ij} = 0$, otherwise. The powers of \mathbf{J} form a cycle with $\mathbf{J}^p = \mathbf{I}$, where \mathbf{I} is the $p \times p$ identity matrix. An argument similar to that in Section 2 leads to $\text{Var}(x_h) = \sigma^2 \mathbf{D}^2$ for all h , and, in general, $\text{Cov}(x_h, x_{h-r}) = \sigma^2 \mathbf{D} \mathbf{P}_r \mathbf{J}^r \mathbf{D}$, for $r = 0, 1, 2, \dots$, and all h .

Finally, define the matrix Z as $\sum_{n=1}^{\infty} k^n P_n J^n$. For general p , i , and j , it can be shown that the ij th cell Z_{ij} is an infinite sum of terms:

$$Z_{ij} = \sum_{m=0}^{\infty} k^m \rho_{u,i}, \quad \text{where } u = pm + 1 + \text{mod}_p(p - i + j - 1).$$

Because the ρ values represent correlation coefficients, it follows easily that Z is finite.

Analogous to theorems 1, 2, and 3 proven earlier are theorems 4, 5, and 6 presented below. The former three allow any pattern of design gaps, but apply only to one-level designs. Theorems 4, 5, and 6 do not permit designs gaps.

The proofs of the theorems are similar to those in Section 3 and are not repeated. All results apply to the limiting case where rotation groups have been in sample long enough to eliminate the effect of phasing in the sample. If the $\rho_{r,i}$'s decrease rapidly with r , or if k is relatively small, the "steady-state" arrives within a couple of interviews.

Theorem 4. If the GCE of level is defined as in (13), and the covariance structure of (14) holds, then

$$\begin{aligned} \text{Var}(y_h) = & \sigma^2 \{ a^T D^2 a + k^2 b^T D^2 (b - 2a) \\ & + 2(a - k^2 b)^T D Z D (a - b) \} / (1 - k^2). \end{aligned}$$

Theorem 5. Let $y_h - y_{h-1}$ be the GCE estimator of "month-to-month" change. Then $\text{Var}(y_h - y_{h-1})$ is

- (i) $2\sigma^2 a^T D (I - P_1 J) D a$, if $k = 0$, and
- (ii) $\sigma^2 (a^T D^2 a + k^2 b^T D^2 b - 2ka^T D P_1 J D b) / k - (1 - k)^2 \text{Var}(y_h) / k$, if $0 < k < 1$.

Theorem 6. Define $S_{h,t}$ as in (9), the sum of the GCE's for the last t periods. Then $S_{h,t}$, $S_{h,t} - S_{h-t,t}$, and $y_h - y_{h-t}$ can again be written as $\sum_{i=0}^{\infty} v_i^T x_{h-i}$, where the vectors v_0, v_1, v_2, \dots are found in (10). For these sets of vectors,

$$\text{Var} \left(\sum_{i=0}^{\infty} v_i^T x_{h-i} \right) = \sigma^2 \left\{ \sum_{i=0}^{\infty} v_i^T D^2 v_i + 2 \sum_{i=0}^{\infty} v_i^T \sum_{n=1}^{\infty} D P_n J^n D v_{i+n} \right\}; \quad (16)$$

the sums in (16) converge.

ACKNOWLEDGMENTS

The author wishes to thank Lynn Weidman and Larry Ernst for checking the text and proofs. Lynn graciously read through several drafts, and offered many helpful suggestions to improve the presentation of the paper. The comments and suggestions of an associate editor and three referees are also noted and greatly appreciated.

REFERENCES

- BAILAR, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BREAU, P., and ERNST, L. R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.
- GURNEY, M., and DALY, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 242-257.
- HANSON, R. H. (1978). The Current Population Survey: Design and Methodology. Technical Paper 40, U.S. Bureau of the Census, Washington, D.C.
- HUANG, E. T., and ERNST, L. R. (1981). Comparison of an alternative estimator to the current composite estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 303-308.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 403-408.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1984). An Overview of the Survey of Income and Program Participation. SIPP Working Paper Series, No. 8401, U.S. Bureau of the Census, Washington, D.C.
- RAO, J. N. K., and GRAHAM, J. E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- WOLTER, K. M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- WOODRUFF, R. S. (1963). The use of rotating samples in the Census Bureau's monthly surveys. *Journal of the American Statistical Association*, 58, 454-467.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(\cdot)" and "log(\cdot)", etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

- 1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
- 2. **Résumé**
 - Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. **Bibliographie**

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

- KUMAR, S., et LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 403-408.
- NELSON, D., McMILLEN, D., et KASPRZYK, D. (1984). An Overview of the Survey of Income and Program Participation. SIPP Working Paper Series, No. 8401, U.S. Bureau of the Census, Washington, D.C.
- RAO, J. N. K., et GRAHAM, J. E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- WOLTER, K. M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- WOODRUFF, R. S. (1963). The use of rotating samples in the Census Bureau's monthly surveys. *Journal of the American Statistical Association*, 58, 454-467.

Théorème 4. Si l'ECG de niveau est défini comme en (13) et que la structure de covariances définie en (14) est vérifiée, alors,

$$\text{Var}(y_h) = \sigma^2 \{ a^T D^2 a + k^2 b^T D^2 (b - 2a) + 2(a - k^2 b)^T D Z D (a - b) \} / (1 - k^2).$$

Théorème 5. Soit $y_h - y_{h-1}$ de la variation l'ECG "d'un mois à l'autre". Alors, $\text{Var}(y_h - y_{h-1})$ est

$$(i) \quad 2\sigma^2 a^T D (I - P_1 J) D a, \quad \text{si } k = 0, \quad \text{et}$$

$$(iii) \quad \sigma^2 (a^T D^2 a + k^2 b^T D^2 b - 2ka^T D P_1 J D b) / k - (1 - k)^2 \text{Var}(y_h) / k, \quad \text{si } 0 < k < 1.$$

Théorème 6. Définissons $S_{h,t}$ comme dans l'équation (9), soit la somme des ECG relatifs aux t dernières périodes. Alors, $S_{h,t}, S_{h,t-1}, \dots, S_{h-t,t},$ et $y_h - y_{h-t}$ peuvent être exprimées de nouveau par $\sum_{i=0}^\infty v_i^T x_{h-t-i},$ où les vecteurs v_0, v_1, v_2, \dots sont définis en (10). Pour ces séries de vecteurs,

$$\text{Var} \left(\sum_{i=0}^\infty v_i^T x_{h-t-i} \right) = \sigma^2 \left\{ \sum_{i=0}^\infty v_i^T D^2 v_i + 2 \sum_{i=0}^\infty \sum_{n=1}^\infty v_i^T D P^n J^n D v_{i+n} \right\}; \quad (16)$$

les sommes de l'équation (16) convergent.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à Lynn Weidman et à Larry Ernst pour avoir relu le texte et corrigé les épreuves. Lynn s'est fait un plaisir de passer en revue plusieurs versions préliminaires et a fait de nombreuses suggestions visant à améliorer la présentation de cet article. Les commentaires et suggestions d'un rédacteur associé et de trois arbitres ont été grandement appréciés.

BIBLIOGRAPHIE

BAILLAR, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BREAU, P., et ERNST, L. R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

GURNEY, M., et DALY, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 242-257.

HANSON, R. H. (1978). The Current Population Survey: Design and Methodology. Technical Paper 40, U.S. Bureau of the Census, Washington, D.C.

HUANG, E. T., et ERNST, L. R. (1981). Comparison of an alternative estimator to the current composite estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 303-308.

$$i) \text{ Var}(x_{h,i}) = d_i^2 \sigma^2 \text{ pour tous } h \text{ et tous } i, \text{ où } d_i > 0;$$

$$ii) \text{ Cov}(x_{h,i}, x_{h,j}) = 0 \text{ pour } i \neq j;$$

$$iii) \text{ Pour } r \geq 0: \text{ Cov}(x_{h,i}, x_{h-r,j}) = \rho_{r,i} d_i d_j \sigma^2, \text{ si les deux valeurs } x \text{ concernent le même groupe à } r \text{ mois d'intervalle, ou } 0 \text{ dans les autres cas. Posons}$$

$$(14)$$

Il se pourrait bien que $d_1 \leq d_2 \leq \dots \leq d_p$ si la variabilité de réponse s'accroît avec la période de remémoration. L'indice r du coefficient de corrélation $\rho_{r,i}$ représente la période qui sépare les mois auxquels se rapportent les estimations $x_{h,i}$ et $x_{h-r,j}$. L'indice i signifie que l'estimation pour le mois h a été établie par suite d'une interview réalisée i mois plus tard. Pour des valeurs particulières de h, r et i , il existe une seule valeur de $j, 1 \leq j \leq p$ pour laquelle les estimations $x_{h,i}$ et $x_{h-r,j}$ ont trait au même panel et la $\text{Cov}(x_{h,i}, x_{h-r,j})$ est non nulle. (Cette valeur est $j = \text{mod}_p(i + r - 1) + 1$, où $\text{mod}_p(n)$ est la valeur de l'entier n , modulo p .) Autrement, la covariance est nulle. Dans certains cas, il peut être approprié de remplacer $\rho_{r,1}, \dots, \rho_{r,p}$ par une seule valeur ρ_r .

Aucune hypothèse n'est posée au sujet des biais. Outre l'effet de la période de remémoration sur la variance des estimations des groupes de renouvellement, comme nous le supposons en (14), il pourrait exister un biais dû à cette période. Nous pourrions aussi observer un biais dû au nombre de mois passés dans l'échantillon, biais qui dépend du nombre de fois qu'un répondant a été interviewé (Bailar 1975). Bien qu'il ne soit pas nécessaire de connaître ces biais pour résoudre les formules de variance définies dans cette section, ils peuvent représenter un élément non négligeable de l'erreur quadratique moyenne.

Définissons les matrices D, P_r et J de dimensions $p \times p$ de la façon suivante. Soient D et P_r , pour $r \geq 0$, des matrices diagonales ayant pour éléments diagonaux d_1, \dots, d_p et $\rho_{r,1}, \dots, \rho_{r,p}$, respectivement. Définissons J comme suit: $J_{i,i+1} = 1$ pour $i = 1, 2, \dots, p - 1$; $J_{p,1} = 1$ et $J_{ij} = 0$ dans les autres cas. Un argument semblable à celui présenté dans la section 2 nous permet d'affirmer que $\text{Var}(x_h) = \sigma^2 D^{\cup} D^2$ pour h et, de façon générale, que $\text{Cov}(x_h, x_{h-r}) = \sigma^2 D P_r J^r D$, pour $r = 0, 1, 2, \dots$, et tous h .

Enfin, définissons la matrice Z comme $\sum_{n=1}^{\infty} k^n P_n J^n$. Pour les valeurs p, i et j en général, on peut montrer que la fréquence Z_{ij} de la case ij est une somme infinie de termes:

$$Z_{ij} = \sum_{n=0}^{\infty} k^n p_{n,i}, \text{ où } n = pm + 1 + \text{mod}_p(p - i + j - 1).$$

Comme les valeurs p représentent des coefficients de corrélation, il est facile de déduire que Z est finie.

Les théorèmes 4, 5 et 6 énoncés ci-dessous sont analogues aux théorèmes 1, 2 et 3 qui ont été démontrés précédemment. Ces derniers permettent n'importe quel genre d'intervalles de plan mais concernent uniquement les plans à un niveau. Les théorèmes 4, 5 et 6 ne permettent pas d'intervalle de plan.

Les démonstrations des théorèmes sont semblables à celles de la section 3 et ne sont pas reprises ici. Tous les résultats s'appliquent au cas limite où les groupes de renouvellement font partie de l'échantillon depuis assez longtemps pour qu'il n'y ait plus d'effet d'échelonnement dans l'échantillon. Si les $\rho_{r,i}$ diminuent rapidement avec r , ou si k est relativement petit, l'état stationnaire est atteint en deux interviews.

Tableau 1
Présentation des estimations selon un plan longitudinal à quatre niveaux

MOIS	GROUPES DE RENOUVELLEMENT			
↑	→			
1	1			
2	2			
3	3			
4	4			

Nota: $x_{h,i}$ représente l'estimation de la valeur mensuelle de la caractéristique pour le mois h , établie à l'aide des données du groupe de renouvellement interviewé au mois $h + i$. Les interviews débutent au mois 5. Les traits horizontaux (coulures) servent à séparer les estimations tirées d'interviews différentes

Posons $x_h = (x_{h,1}, x_{h,2}, \dots, x_{h,p})^T$ comme le vecteur formé des p estimations établies pour le mois h à l'aide des données obtenues des p groupes dans des interviews différentes. Notons que $x_{h,p}, x_{h+1,p-1}, \dots, x_{h+p-1,1}$ sont les estimations relatives à p mois différents, établies à l'aide des données recueillies auprès d'un groupe dans une seule interview (au mois $h + p$). Comme dans les sections 2 et 3, l'estimateur composite généralisé de niveau mensuel est défini

(13) $y_h = \sum a_i x_{h,i} - k \sum b_i x_{h-1,i} + k y_{h-1}$

où l'indice des sommations prend les valeurs de 1 à p . En définissant a et b comme $(a_1, \dots, a_p)^T$ et $(b_1, \dots, b_p)^T$ respectivement, nous pouvons de nouveau formuler l'ECCG comme suit:

$$y_h = a^T x_h - k b^T x_{h-1} + k y_{h-1}$$

On suppose que la structure de covariances des estimations mensuelles des groupes de renouvellement est stationnaire dans le temps. En revanche, le délai qui sépare le mois de référence h de l'interview correspondante (au mois $h + i$ peut influencer sur la variabilité des réponses, $x_{h,i}$. Pour $i = 1, \dots, p$, soit d_i^2 la variabilité de réponse exprimée comme une fonction du délai qui sépare le mois de référence de l'interview. Nous posons par hypothèse la structure de covariances suivante:

Nous nous servons ensuite des équations (12) et (5) pour déterminer les trois séries de vecteurs v_i . La démonstration de la partie b) est semblable à celle du théorème 1, une fois que l'on a constaté que les v_i définis en (a) sont des vecteurs à l'image du plan étant donné qu'ils sont des combinaisons linéaires de a et de $a - b$. Afin de démontrer la convergence, soulignons que, pour les trois séries de vecteurs définis en (a), v_i est proportionnel à $k'(a - b)$ pour une valeur i suffisamment élevée. Il existe une constante $\lambda > 0$ telle que, pour $i \geq 2t$ et chaque élément j , $|v_{ij}| \leq k'\lambda$. Or, nous avons vu que $|p_j| \leq 1$ et que chaque ligne de L^n contient tout au plus un élément non nul (égal à 1); par conséquent, la somme finie de l'équation (11) est considérée comme un vecteur $M \times 1$ dont chacun des éléments est borné supérieurement en valeur absolue par $k'(M - 1)\lambda$. Donc, la sommation double converge puisque k'^2 converge.

4. PLANS À PLUSIEURS NIVEAUX

Bien que les résultats obtenus dans les sections 2 et 3 s'appliquent à tous les plans de renouvellement équilibrés à un niveau, on constate que de nombreuses enquêtes reposent sur des plans à plusieurs niveaux. C'est le cas notamment de la Survey of Income and Program Participation (SIPP); dans cette enquête, un groupe de renouvellement sur quatre est interviewé à chaque mois et les répondants doivent alors fournir des renseignements à propos des quatre mois précédents. Quoique le plan puisse être modifié à tout moment, le premier groupe de renouvellement est interviewé aux mois de février, juin, octobre, février, etc. jusqu'à concurrence de huit interviews. Le second groupe est interviewé en mars, juillet, etc. Les deux autres groupes sont soumis aux mêmes règles de sondage sauf que dans un cas, la série d'interviews débute en avril et dans l'autre, elle débute en mai. Considérés en bloc, les quatre groupes de renouvellement forment ce qu'on appelle un panel; celui-ci couvre une période d'environ deux ans et demi. à chaque année, on introduit un nouveau panel. Par exemple, le panel de 1986 a duré de 1986 à 1988 tandis que celui de 1987 a couvert la période 1987-1989. On prend soin de ne pas combiner des données de panels différents même s'il s'agit de données de la même période. Pour plus de détails sur le plan de la SIPP, voir Nelson, McMillen et Kasprzyk (1984). Dans le cas des plans à un niveau, le schéma des interviews et des intervalles de plan – période où il n'y a aucune interview – pour un groupe de renouvellement peut être très varié, pourvu qu'il s'agisse d'un plan équilibré. Dans le cas des plans à plusieurs niveaux toutefois, l'existence d'intervalles de plan peut causer des problèmes au point de vue de l'effort de mémoire. S'il doit se reporter plusieurs mois en arrière, un répondant pourrait avoir de la difficulté à situer un événement à la bonne période. Les intervalles de plan ne peuvent qu'ajouter à la confusion. Pour cette raison et parce que les plans à plusieurs niveaux qui renferment des intervalles de plan sont rares dans la pratique, nous n'allons considérer ici que des plans selon lesquels l'échantillon est formé de p groupes de renouvellement, ii) les groupes sont interviewés successivement à tous les p "mois" et iii) la période de référence correspond aux p mois précédents. De nombreuses enquêtes comme la National Crime Survey, qui est réalisée sous l'égide du U.S. Bureau of Justice Statistics, ont un plan de renouvellement plus compliqué que ceux analysés ici. Il est normal que les formules de variance pour estimateurs composites soient plus complexes dans ces circonstances.

Interviewer un groupe de renouvellement signifie que l'on recueille simultanément des données auprès de toutes les unités de ce groupe à un mois précis. Pour une caractéristique particulière qui doit être estimée, définissons x_{hi} comme l'estimation de la valeur mensuelle de cette caractéristique pour le mois h , établie à l'aide des données du groupe de renouvellement interviewé au mois $h + i$, où $i = 1, \dots, p$. L'indice i désigne la "période de renouvellement", c'est-à-dire la période écoulée entre le mois de référence et le moment de l'interview. Le tableau 1 donne les estimations x_{hi} pour un plan à quatre niveaux et à quatre groupes de renouvellement. Les traits horizontaux qui figurent dans le diagramme servent à séparer les estimations tirées d'interviews différentes. Ces traits sont appelés "coutures" dans la SIPP.

on pourra isoler le terme de covariance pour une application ultérieure. Enfin, $y_h - y_{h-1} = W_h - (1 - k)y_{h-1}$. En nous servant des équations (7) et (8) dans le calcul de la variance de cet écart, nous obtenons l'expression voulue.

Souvent, on sera plus intéressé par le niveau moyen pour une période donnée (par ex.: un trimestre ou une année), la différence entre ces moyennes d'une "année" à l'autre ou encore la différence de niveau "mensuel" pour deux mois situés à un an d'intervalle. Désignons par $S_{h,t}$ la somme des ECG relatifs aux t derniers mois:

$$(9) \quad S_{h,t} = y_h + y_{h-1} + \dots + y_{h-t+1}, \quad t \geq 1.$$

Les valeurs les plus courantes de t sont trois, quatre et douze. Si l'on veut une moyenne plutôt qu'une somme, il suffira de diviser $S_{h,t}$ par t .

Théorème 3:

a) Les expressions $S_{h,t}$, $S_{h,t} - S_{h-t,t}$, et $y_h - y_{h-t}$ peuvent être écrites sous la forme $\sum_{i=0}^{\infty} v_i^T x_{h-1}$, où

$$\begin{aligned} & \text{i) pour } S_{h,t}, v_i = : \\ & a + [(k - k_{t+1}) / (1 - k)](a - b), \text{ pour } i = 0, 1, \dots, t - 1, \\ & [k_{t-t}^t (k - k_{t+1}) / (1 - k)](a - b), \text{ pour } i = t, t + 1, t + 2, \dots; \\ & \text{ii) pour } S_{h,t} - S_{h-t,t}, v_i = : \\ & a + [(k - k_{t+1}) / (1 - k)](a - b), \text{ pour } i = 0, 1, \dots, t - 1, \\ & [(2k_{t-t+1} - k - k_{t+1}) / (1 - k)](a - b) - a, \text{ pour } i = t, t + 1, \dots, 2t - 1, \\ & - [k_{t-2t+1}^t (1 - k_{t+1}) / (1 - k)](a - b), \text{ pour } i = 2t, 2t + 1, \dots; \\ & \text{iii) pour } y_h - y_{h-t}, v_0 = a, v_i = k^i(a - b) - a, \text{ et } v_i = : \\ & k^i(a - b), \text{ pour } i = 1, 2, \dots, t - 1, \\ & - k_{t-t}^t (1 - k^t)(a - b), \text{ pour } i = t + 1, t + 2, \dots; \end{aligned}$$

b) Pour les séries de vecteurs v_0, v_1, v_2, \dots définis en a),

$$(11) \quad \text{Var} \left(\sum_{i=0}^{\infty} v_i^T x_{h-i} \right) = \sigma^2 \left\{ \sum_{i=0}^{\infty} v_i^T v_i + 2 \sum_{i=0}^{\infty} \sum_{n=1}^{\infty} \rho^n T_n^T v_{i+n} \right\};$$

les sommes de l'équation (11) convergent.

Démonstration du théorème 3. Partie a): en incluant successivement les termes y_h à y_{h-t+1} et en appliquant l'équation (5) à y_{h-t} , nous obtenons

$$\begin{aligned} S_{h,t} &= a^T(x_h + x_{h-1} + \dots + x_{h-t+1}) + k(a - b)^T x_{h-1} \\ &+ (k + k^2)(a - b)^T x_{h-2} + \dots \\ &+ (k + k^2 + \dots + k^{t-1})(a - b)^T x_{h-t+1} \\ &+ (k + k^2 + \dots + k^t)(a - b)^T \sum_{j=t}^{\infty} k_{j-t}^j x_{h-j}. \end{aligned}$$

(12)

$$\begin{aligned}
 & + 2a^T \sum_{i=1}^{\infty} k_i \sigma_2 \rho_i R L_i R(a-b) \\
 & + 2(a-b)^T \sum_{i=1}^{\infty} \sum_{j=i+1}^{\infty} k_{i+j} \sigma_2 \rho_j^{-i} R L_j^{-i} R(a-b) \\
 & = \sigma_2 \{ a^T R a + (a-b)^T R(a-b) k^2 / (1-k^2) \\
 & + 2a^T R \left(\sum_{i=1}^{\infty} k_i \rho_i L_i \right) R(a-b) \\
 & + 2(a-b)^T R \left(\sum_{i=1}^{\infty} k_{2i} \left[\sum_{j=i+1}^{\infty} k_{j-i} \rho_j^{-i} L_j^{-i} \right] R(a-b) \right) \}. \quad (6)
 \end{aligned}$$

Comme a et $a-b$ sont des vecteurs à l'image du plan, $a^T R = a^T$, $(a-b)^T R = (a-b)^T$, et $R(a-b) = (a-b)$. La somme $\sum_{i=1}^{\infty} k_i \rho_i L_i$ est considérée comme la matrice \bar{Q} : l'élément ij de cette matrice est $k_{i-j} \rho_j^{-i} L_j^{-i}$, si $1 \leq j < i \leq M$, et 0 si ce n'est pas le cas. Par un changement de variables, nous pouvons voir que la somme entre crochets est aussi la matrice \bar{Q} . Nous pouvons alors reformuler l'expression (6) comme suit:

$$\sigma_2 \{ a^T a + (a-b)^T (a-b) k^2 / (1-k^2) + 2a^T \bar{Q}(a-b) + 2(a-b)^T \bar{Q}(a-b) k^2 \}.$$

Par une simple transformation mathématique, nous obtenons l'équation (4).

Théorème 2. Soit $y_h - y_{h-1}$ l'ECG de la variation d'un "mois à l'autre". Alors, $\text{Var}(y_h - y_{h-1})$ est

$$\begin{aligned}
 & \text{i) } 2\sigma_2^T (I - \rho_1 L) a, \text{ si } k = 0, \text{ et} \\
 & \text{ii) } \sigma_2^T (a^T a + k^2 b^T b - 2k \rho_1 a^T L b) / k, \text{ si } 0 < k < 1.
 \end{aligned}$$

Démonstration du théorème 2:

i) Si $k = 0$, $y_h = a^T x_h$. D'après (3), la variance de $a^T x_h - a^T x_{h-1}$ est

$$2a^T \sigma_2^T R a - 2a^T \sigma_2^T \rho_1 R L R a = 2\sigma_2^T a^T (I - \rho_1 L) a.$$

ii) Si $0 < k < 1$, définissons W_h comme $a^T x_h - k b^T x_{h-1}$. D'après les résultats antérieurs, nous constatons facilement que

$$\text{Var}(W_h) = \sigma_2^T \{ a^T a + k^2 b^T b - 2k \rho_1 a^T L b \}. \quad (7)$$

D'après (1a), $y_h = W_h + k y_{h-1}$. Alors

$$\text{Var}(y_h) = \text{Var}(W_h) + k^2 \text{Var}(y_{h-1}) + 2k \text{Cov}(W_h, y_{h-1}); \quad (8)$$

Nous supposons que les données affichent une structure de covariances stationnaire:

- i) $\text{Var}(x_{h,i}) = \sigma^2$ pour tous h et i ;
 ii) $\text{Cov}(x_{h,i}, x_{h,j}) = 0$ pour $i \neq j$, c.-à-d. qu'il n'existe aucune corrélation entre deux groupes de renouvellement dans le même mois;
 iii) $\text{Cov}(x_{h,i}, x_{s,j}) = \rho^{|h-s|} \sigma^2$, si les deux observations x 's concernent le même groupe de renouvellement à $|h-s|$ mois d'intervalle, ou 0, dans les autres cas. Soit $\rho_0 = 1$.

D'après les équations i) et ii), il est clair que $\text{Var}(x_h) = \sigma^2 R$, pour tous h . L'équation iii) implique que $\text{Cov}(x_h, x_{h-1}) = \sigma^2 \rho_1 R L R$ parce que la matrice L , dont la sous-diagonale est formée de uns, "représente" le décalage d'un mois entre les valeurs x_h et x_{h-1} et que la prémultiplication (postmultiplication) par R a pour effet d'introduire des valeurs nulles qui correspondent aux valeurs nulles dans x_h (x_{h-1}) (mois où le groupe de renouvellement ne fait pas partie de l'échantillon). Nous constatons facilement que $(L^T)_{ij} = 1$ si $i - j = r \geq 0$ et $1 \leq j, i \leq M$; définissons L^0 comme la matrice unité. En appliquant le même raisonnement que ci-dessus, nous obtenons $\text{Cov}(x_h, x_{h-2}) = \sigma^2 \rho_2 R L^2$. D'une manière générale,

$$(3) \quad \text{Cov}(x_h, x_{h-r}) = \sigma^2 \rho_r R L^r R, \text{ pour } r = 0, 1, 2, \dots, \text{ et tous } h.$$

Pour $r \geq M$, $L^r = 0$, et $\text{Cov}(x_h, x_{h-r}) = 0$.

Pour les théorèmes qui suivent, définissons la matrice \tilde{Q} de dimension $M \times M$ comme suit: $\tilde{Q}_{ij} = k_{i-j} \rho_{i-j}$ si $1 \leq j < i \leq M$ et 0 si ce n'est pas le cas. Enfin, soit I la matrice unité $M \times M$.

3. PLANS À UN NIVEAU: THÉORÈMES ET DÉMONSTRATIONS

Nous allons maintenant énoncer trois théorèmes et en faire la démonstration.

Théorème 1. Si l'EBC de niveau est défini comme en (1) et que la structure de covariances définie en (2) est juste, alors

$$(4) \quad \text{Var}(y_h) = \sigma^2 \{ a^T a + k^2 b^T (b - 2a) + 2(a - k^2 b)^T \tilde{Q} (a - b) \} / (1 - k^2).$$

Il convient de souligner que lorsqu'on utilise une moyenne non pondérée des estimations des m groupes de renouvellement du mois courant, $k = 0$, $\tilde{Q} = 0$, et $a_i = 1/m$, pour $i = 1, \dots, m$. Par conséquent, $\text{Var}(y_h) = \sigma^2/m$, comme prévu.

Démonstration du théorème 1. En appliquant un processus de substitution récursif à l'équation (1a), nous obtenons

$$(5) \quad y_h = a^T x_h + (a - b)^T \sum_{i=1}^{\infty} k^i x_{h-i}.$$

D'après (3), la variance de cette somme est

$$\text{Var}(y_h) = a^T \sigma^2 R a + (a - b)^T \sum_{i=1}^{\infty} k_{2i} \sigma^2 R (a - b)$$

Le vecteur w , de dimension $M \times 1$, comme suit. Définissons le i -ième élément de w comme étant égal à 0 si $i \in T_0$. Nous comptons ainsi $M - m$ positions du vecteur w . Dans un second temps, nous comptons les m positions qui restent avec les valeurs w_1, \dots, w_m suivant l'ordre logique. Nous obtenons ainsi un vecteur dit "à l'image du plan". Par exemple, dans le cas d'un plan de renouvellement de type 4-8-4, $T_0 = \{5, 6, \dots, 12\}$ et $w^T = (w_1, w_2, w_3, w_4, 0, 0, 0, 0, 0, 0, 0, 0, w_5, w_6, w_7, w_8)$.

prémultiplication par R a pour effet d'«éliminer» (de ramener à 0) les lignes de V identifiées par des éléments de T_0 . Si les colonnes de V sont déjà à l'image du plan, alors $RV = V$. De même, pour n'importe quelle matrice U de dimension $p \times M$, la postmultiplication par R a pour effet d'«éliminer» les colonnes de U qui sont identifiées par des éléments de T_0 . Si les colonnes de U ne sont pas dans le plan, il y a une partie non nulle de U qui est envoyée vers T_0 et donc éliminée.

Si nous considérons maintenant les données proprement dites, soit x_{hi} , l'estimation de la valeur "mensuelle", d'une caractéristique à mesurer, tirée du groupe de renouvellement qui a fait partie de l'échantillon pour la i -ième fois au mois h , où $i = 1, \dots, m$. Brau et Ernst (1983) ont défini l'estimateur composite généralisé (ECG) de niveau par un processus récursif. Pour ce qui a trait au niveau mensuel, soit:

où $0 \leq k < 1$, et a_i 's et b_i 's peuvent n'importe quelle valeur (y compris des valeurs négatives) à la condition que $\sum_{i=1}^m a_i = 1$ et $\sum_{i=1}^m b_i = 1$. L'estimateur composite courant et l'estimateur composite AK utilisés dans la CPS sont des cas particuliers de l'ECG. Pour plus de détails sur ces estimateurs, voir Hanson (1978), Huang et Ernst (1981) et Kumar et Lee (1983).

Pour faciliter le calcul de la variance, nous allons exprimer l'équation (1) sous forme vectorielle. Soient a et b des vecteurs $M \times 1$ à l'image du plan formés des séries de constantes a_1, \dots, a_m et b_1, \dots, b_m respectivement. De même, pour n'importe quelle valeur de h , les observations $x_{h,1}, \dots, x_{h,m}$ constituent le vecteur x_h , qui est aussi un vecteur $M \times 1$ à l'image du plan. Alors,

$$y_h = x_h^T x - k x_h^{T-1} + k y_{h-1}. \quad (1a)$$

Lorsqu'il cherche à déterminer le plan de renouvellement optimal et l'estimateur composite optimal, l'utilisateur doit décider du niveau de variance le plus acceptable pour les estimateurs de niveau, de variation mensuelle et de moyenne.

Dans la section 4, nous étendons les résultats de notre analyse à certains plans de renouvellement à plusieurs niveaux, notamment à la SIPP. Sauf quelques restrictions mineures – en particulier, l'inexistence d'intervalles de plan – nous enonçons des théorèmes semblables à ceux de la section 3. Comme les calculs sont les mêmes que pour les plans à un niveau, nous ne faisons pas la démonstration des résultats.

2. PLANS À UN NIVEAU: NOTATION ET DÉFINITIONS

Bien que les formes de plans de renouvellement soient infiniment nombreuses, nous nous limitons dans les sections 2 et 3 à l'étude d'une seule forme. À chaque passage de l'enquête, un nouveau groupe de renouvellement est introduit dans l'échantillon et il y demeure et en est exclu selon les mêmes règles de durée que les groupes qui l'ont précédé. En outre, l'interview porte uniquement sur la période courante, peu importe que les répondants aient fait partie ou non de l'échantillon à la période précédente. C'est ce qu'on appelle un plan de renouvellement "équilibré à un niveau". On parle d'un plan "équilibré" parce que le nombre de groupes présents dans l'échantillon à n'importe quelle période est égal au nombre total de périodes qu'un groupe passe dans l'échantillon.

Le plan de renouvellement prévu dans l'EPA répond à ces conditions. Un nouveau groupe est introduit dans l'échantillon à chaque mois et y demeure cinq autres mois. Dans sa version actuelle, la CPS prévoit un plan de renouvellement de type 4-8-4 (4 mois dans l'échantillon, 8 mois hors de l'échantillon, 4 mois dans l'échantillon). Avant juillet 1953, la CPS utilisait un plan non équilibré, en vertu duquel les groupes de renouvellement, au nombre de cinq, étaient introduits successivement dans l'échantillon à raison d'un par mois. Le sixième mois, *aucun nouveau groupe n'était introduit dans l'échantillon*. Les groupes se succédaient selon ce mode de renouvellement et étaient supprimés de l'échantillon au bout de six mois.

Une des difficultés que posait le plan de renouvellement de la CPS avant 1953 était l'introduction d'un biais dû au nombre de mois passés dans l'échantillon, souvent appelé biais de renouvellement. Mais ce qui nous préoccupe davantage ici est l'irrégularité avec laquelle se succédaient les groupes de renouvellement. La variance d'un estimateur composite dépend du nombre de mois écoulés depuis l'introduction de chaque groupe dans l'échantillon et de la structure de covariances pour des groupes identiques à des mois différents. Si l'ordre de succession des groupes dans l'échantillon varie d'un mois à l'autre, la formule de la variance de l'estimateur variera également. Avec un plan équilibré et une structure de covariances stationnaire, nous pouvons définir des formules générales.

Pour des raisons de concision et parce que le mois est l'unité de temps fondamentale dans la plupart des enquêtes des organismes d'Etat, nous allons nous servir du mot "mois" dans cet article pour désigner la période où ont lieu les interviews. Néanmoins, les résultats consignés dans cette section et la suivante s'appliquent à n'importe quelle période de temps, pourvu qu'il s'agisse d'un plan de renouvellement équilibré à un niveau.

Nous allons maintenant présenter la notation et définir certains vecteurs. Supposons que chaque groupe de renouvellement doit subir en tout m interviews sur une période de M mois. Autrement dit, chaque groupe se trouve hors de l'échantillon pendant $M - m$ mois entre le moment où il y est introduit pour la première fois et le moment où il en est supprimé définitivement. Le plan équilibré fait qu'il y a toujours m groupes dans l'échantillon, peu importe le mois.

Nous définissons l'ensemble T_0 comme suit. Considérons un groupe de renouvellement quelconque: T_0 désignera l'ensemble des "mois" où ce groupe ne fait pas partie de l'échantillon. On identifiera par le chiffre 1 le mois où ce groupe est interviewé pour la première fois,

estimateurs ou les frais liés au processus d'échantillonnage sans imposer un fardeau de réponse excessif aux participants. Dans la CPS et l'EPA, les taux de chevauchement des échantillons d'un mois à l'autre sont de 75 et de 83% respectivement. Pour approfondir ces questions, voir Woodruff (1963), Rao et Graham (1964) ou Wolter (1979).

Certains estimateurs utilisés avec les plans de renouvellement sont "composites" par définition. Ils intègrent des estimations du mois courant et des mois antérieurs pour un groupe de renouvellement donné, ce qui permet de tirer profit de l'échantillonnage répété. Ces estimations sont fondées en un estimateur final.

Bien que l'on puisse réduire la variance des estimateurs composites par une intégration judicieuse des données, il peut être difficile de la calculer à cause de la corrélation qui peut exister entre les groupes répétés. Dans cet article, nous définissons des formules simples pour la variance d'estimateurs de niveau et de variation pour un plan de renouvellement général assujéti à des contraintes particulières. Les expressions s'appliquent à une importante catégorie d'estimateurs appelés estimateurs composites généralisés (Breau et Ernst 1983).

Ces formules seront utiles dans la mesure où nous pourrions estimer la corrélation entre les estimations du même groupe de renouvellement pour des périodes différentes et dans la mesure où cette corrélation sera suffisamment élevée pour justifier l'application de l'estimation composite. Dans les enquêtes permanentes, les données d'échantillon antérieures permettent habituellement d'estimer cette corrélation. Pour les caractéristiques qui ont trait au revenu du ménage ou à la population active, la corrélation est, en règle générale, moyennement élevée. Toutefois pour d'autres caractéristiques, comme le taux de criminalité, la corrélation entre les périodes peut ne pas être suffisamment élevée pour justifier l'utilisation de l'estimation composite. De toutes les enquêtes mentionnées dans cet article, seule la CPS utilise ordinairement un estimateur composite.

Dans l'exposé qui suit, nous examinons deux types d'enquêtes différents. Dans des enquêtes comme la CPS et l'EPA, les participants ne fournissent des données que pour le mois courant. C'est ce qu'on appelle des enquêtes à un niveau. Par ailleurs, le U.S. Census Bureau réalise la Survey of Income and Program Participation (SIPP) afin de recueillir des données sur le niveau et les sources de revenu, la participation aux programmes et d'autres points pertinents. À chaque interview de cette enquête, les répondants doivent fournir des données qui ont rapport aux quatre mois précédents. Un nouveau groupe est interviewé à chaque mois. C'est pourquoi le plan de la SIPP est appelé plan à plusieurs niveaux. Wolter (1979) parle de plans à un ou à plusieurs niveaux pour indiquer le nombre de périodes pour lesquelles des données sont recueillies au cours d'une interview.

Il convient de faire une autre distinction entre ces deux types d'enquêtes. Appelons "intervalles de plan" une période entre deux interviews qui n'est jamais visée par aucune interview. Tandis que l'EPA ne compte aucun intervalle de ce genre, la CPS en compte un de huit mois. En ce qui concerne les plans à un niveau, le mode de répartition des interviews et des intervalles de plan *ne change rien* aux calculs et aux résultats pertinents. Les formules proposées s'appliquent donc non seulement aux plans de la CPS et de l'EPA mais à d'autres plans qui sont pris en considération.

Pour des raisons que nous verrons plus loin, les intervalles de plan se retrouvent rarement en pratique dans les plans de renouvellement à plusieurs niveaux. La SIPP ne fait pas exception à cette règle. Par conséquent, les plans à plusieurs niveaux que nous étudierons dans cet article ne comptent pas d'intervalles de plan.

Les plans à un niveau font l'objet des sections 2 et 3. Dans la section 2, nous définissons l'estimateur composite généralisé et exposons la notation utilisée. Cette section comprend aussi des définitions et des hypothèses relatives à la covariance. La section 3 contient les éléments principaux – théorèmes 1 à 3. Nous y définissons des variances d'estimateurs de niveau et de variation de niveau. Des formules sont déterminées pour une période précise (par ex.: mois) ou une combinaison de périodes (par ex.: trimestre ou année). Elles s'appliquent aux plans à un niveau, quels que soient le mode de distribution des interviews et les intervalles de plan.

Formules de variance pour estimateurs composites dans les plans de renouvellement

PATRICK J. CANTWELL¹

RÉSUMÉ

Dans de nombreuses enquêtes réalisées par l'Etat, les répondants sont interviewés à un certain nombre de reprises avant que l'enquête ne prenne fin; on parle alors d'enquête avec plan de renouvellement ou d'échantillonnage répété. On a souvent recouru à des estimateurs composites – qui intègrent des données de la période courante et de périodes antérieures – pour déterminer la valeur d'une caractéristique d'intérêt. Comme l'ont mentionné d'autres auteurs, on peut se servir des estimateurs composites dans un plan de renouvellement afin de réduire la variance des estimateurs de variation de niveau. Dans cet article, nous établissons des formules simples pour les variances d'une catégorie d'estimateurs composites de niveau, de variation de niveau et de niveau moyen. Nous considérons tout d'abord des plans de renouvellement à un niveau, où seul le mois courant fait l'objet de l'interview. Nous établissons des résultats pour des plans de sondage qui prévoient m interviews dans une période de M mois. Nous passons ensuite aux plans à plusieurs niveaux, à chaque mois, un groupe parmi p est interviewé. Les membres de ce groupe répondent alors à des questions qui portent sur les p mois antérieurs. Les résultats obtenus dans les diverses sections de cet article s'appliquent à un très grand nombre d'enquêtes des organismes d'Etat.

MOTS CLÉS: Echantillonnage répété dans les enquêtes; plans équilibrés; variation d'un mois à l'autre; moyenne annuelle.

1. INTRODUCTION

Des plans de renouvellement de toutes sortes sont utilisés dans de nombreuses enquêtes-ménages d'envergne. La Current Population Survey (CPS) est réalisée par le U.S. Bureau of the Census pour le compte du U.S. Bureau of Labor Statistics. De son côté, Statistique Canada dirige l'enquête sur la population active (EPA). Les deux enquêtes produisent des estimations des caractéristiques de la population active, y compris le chômage. Dans chaque cas, les ménages sont interviewés un nombre déterminé de fois avant d'être sortis de l'échantillon. Dans la CPS, chaque ménage est tout d'abord introduit dans l'échantillon pour une période de quatre mois, puis en est retiré pour une période de huit mois et est ensuite réintroduit pour une autre période de quatre mois. Dans l'EPA, un ménage fait partie de l'échantillon pendant six mois consécutifs, puis n'y revient pas.

L'enquête avec plan de renouvellement est un intermédiaire entre l'enquête à échantillon constant, où ce sont toujours les mêmes répondants, et l'enquête qui utilise des échantillons indépendants, où les répondants sont interviewés une seule fois avant d'être supprimés de l'échantillon. Dans le premier cas, le fait qu'il s'agisse toujours du même échantillon d'une période à l'autre peut contribuer à réduire au maximum la variance des estimateurs de variation lorsqu'il existe une corrélation positive entre les observations d'une période à l'autre. Un autre avantage de l'enquête à échantillon constant est la réduction de certains frais liés au processus d'échantillonnage; en effet, certaines dépenses ne se répètent plus une fois l'échantillon constitué. En revanche, le fardeau de réponse des membres d'un échantillon constant peut être excessif. L'enquête avec plan de renouvellement est donc un moyen de réduire la variance des

¹ Patrick J. Cantwell, statisticien, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233, E.-U. Cet article est un compte-rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

5. CONCLUSION

On remarque que les résultats des tests pour le taux de réponse multiple sont similaires à ceux pour la multiplicité, ce qui n'est pas surprenant. En comparant les niveaux critiques pour la question sur l'origine ethnique à ceux pour l'identité ethnique, on voit que ce sont les réponses à la question sur l'identité ethnique qui sont le plus affectées par la différence entre les deux versions de questionnaires.

La principale raison pour laquelle on a eu recours aux tests aléatoires est que l'échantillon de l'Essai modulaire 2 n'était pas probabiliste. Il y a cependant d'autres cas où l'emploi de ces tests est approprié. Par exemple, pour un test t de Student de l'égalité des moyennes, on doit avoir recours à l'hypothèse de normalité et on suppose également qu'il y a égalité des variances. Ces présupposés sont inutiles pour un test aléatoire. Il faut toutefois se rappeler que les conclusions qu'on peut tirer à partir des tests aléatoires s'appliquent à l'échantillon et pas nécessairement à la population entière, à moins qu'il ne s'agisse d'un échantillon aléatoire simple.

BIBLIOGRAPHIE

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- BRADLEY, J.V. (1968). *Distribution-free Statistical Tests*. Englewood Cliffs: Prentice-Hall.
- DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.
- EDGINGTON, E.S. (1987). *Randomization Tests*, (2^e éd.). New York: Marcel Dekker.
- FISHER, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Tableau 3
Niveaux critiques pour le taux de réponse multiple et la multiplicité

Question	Région	Multiple			Multiplicité		
		P_x	P_y	α	R_x	R_y	α
ORIGINE	HALIFAX	0.435	0.536	0.087	1.617	1.914	0.062
ORIGINE	QUÉBEC	0.154	0.043	0.001	1.143	0.986	0.001
ORIGINE	MONTREAL	0.185	0.194	0.612	1.141	1.152	0.585
ORIGINE	TORONTO	0.127	0.122	0.393	1.124	1.125	0.495
ORIGINE	WINNIPEG	0.293	0.307	0.622	1.439	1.398	0.345
ORIGINE	VANCOUVER	0.285	0.296	0.621	1.440	1.392	0.280
IDENTITÉ	HALIFAX	0.220	0.335	0.035	1.244	1.502	0.029
IDENTITÉ	QUÉBEC	0.140	0.016	0.001	1.131	0.959	0.001
IDENTITÉ	MONTREAL	0.159	0.125	0.063	1.075	1.044	0.186
IDENTITÉ	TORONTO	0.186	0.120	0.001	1.154	1.075	0.005
IDENTITÉ	WINNIPEG	0.224	0.195	0.248	1.253	1.208	0.298
IDENTITÉ	VANCOUVER	0.186	0.183	0.457	1.182	1.137	0.202

Tableau 4
Niveaux critiques pour certaines variables

Question	Variable	Région	P_x	P_y	α
ORIGINE	FRANÇAIS	QUÉBEC	0.127	0.897	0.001
ORIGINE	FRANÇAIS	MONTREAL	0.038	0.210	0.001
ORIGINE	BRITANNIQUE	HALIFAX	0.321	0.837	0.001
ORIGINE	BRITANNIQUE	MONTREAL	0.034	0.092	0.002
ORIGINE	BRITANNIQUE	TORONTO	0.085	0.135	0.003
ORIGINE	BRITANNIQUE	WINNIPEG	0.167	0.234	0.054
ORIGINE	BRITANNIQUE	VANCOUVER	0.267	0.325	0.065
IDENTITÉ	FRANÇAIS	QUÉBEC	0.138	0.899	0.001
IDENTITÉ	BRITANNIQUE	HALIFAX	0.153	0.828	0.001
IDENTITÉ	BRITANNIQUE	MONTREAL	0.022	0.117	0.001
IDENTITÉ	BRITANNIQUE	TORONTO	0.050	0.215	0.001
IDENTITÉ	BRITANNIQUE	WINNIPEG	0.074	0.276	0.001
IDENTITÉ	BRITANNIQUE	VANCOUVER	0.104	0.325	0.001
IDENTITÉ	ITALIEN	TORONTO	0.412	0.463	0.060

«IRLANDAIS», «ÉCOSSAIS» ou «ANGLAIS» a été cochée. Par exemple, si on veut faire un test quant à la proportion de gens qui ont répondu «FRANÇAIS», on définit μ_x et μ_y comme étant les espérances de la proportion de questionnaires avec la réponse «FRANÇAIS» pour les versions «X» et «Y». Pour toutes les régions, on teste:

$$H_0: \mu_x = \mu_y$$

versus

$$H_1: \mu_x < \mu_y$$

La méthode des tests aléatoires a été utilisée avec 999 permutations. Les résultats sont donnés dans le tableau 4.

Soient μ_x et μ_y les espérances de la proportion d'individus donnant une réponse multiple pour les versions «X» et «Y» respectivement. Pour toutes les régions, excluant celle de Halifax, on teste:

$$H_0: \mu_x = \mu_y \qquad \text{versus} \qquad H_1: \mu_x > \mu_y.$$

Pour Halifax, on utilise la contre-hypothèse $H_1: \mu_x < \mu_y$. La raison pour laquelle on utilise une contre-hypothèse différente pour Halifax comparativement aux autres régions est que la majorité des ménages qui ont été sélectionnés dans cette région sont composés de personnes de souche britannique et que pour cette raison, on s'attend à avoir plus de réponses multiples avec le questionnaire «Y». Dans le questionnaire de type «Y», «CANADIEN» n'est pas un choix de réponse pour les questions sur l'identité et l'origine ethnique, et comme la plupart des personnes de cette région sont d'origine anglaise, écossaise ou irlandaise, ces gens qui ont reçu un questionnaire «Y» ont coché une ou plusieurs de ces cases. Pour ce qui est des ménages qui ont reçu un questionnaire «X», les personnes de ces ménages peuvent tout simplement cocher la case «CANADIEN».

Le niveau critique, α , est calculé de la manière suivante: pour la région de Halifax, étant donné qu'on rejette H_0 si la proportion chez les «X» est significativement inférieure à celle observée chez les «Y», le niveau critique est $RANGP^{x-y}/1000$. Par contre, pour toutes les autres régions, étant donné qu'on rejette H_0 si la proportion chez les «X» est significativement supérieure à celle observée chez les «Y», le niveau critique est $(1001-RANGP^{x-y})/1000$. Les résultats sont présentés dans le tableau 3.

La méthode des tests aléatoires a également été utilisée pour tester la multiplicité, c'est-à-dire le nombre de cases cochées par le répondant, pour les questions sur l'origine et l'identité ethnique dans chacune des régions. Cette fois-ci, au lieu de parler de proportions (P_x, P_y), on parle de ratios (R_x, R_y). Le ratio $R_x (R_y)$ est le nombre moyen de cases cochées par répondant pour une question des questionnaires «X» «Y». Par la suite, la méthode demeure la même et au lieu de parler de $RANGP^{x-y}$, on parle de $RANGR^{x-y}$ et la statistique S se définit maintenant comme étant $R_x - R_y$. Cependant, étant donné la plus grande variabilité des statistiques S pour les tests quant à la multiplicité, on a généré un échantillon aléatoire de 1999 permutations au lieu de 999.

Soient F et G les fonctions de répartition du nombre de cases cochées pour un questionnaire «X» et un questionnaire «Y» respectivement. Pour toutes les régions, excluant celle de Halifax, on teste l'hypothèse:

$$H_0: F = G \qquad \text{versus} \qquad H_1: F(z) \leq G(z) \text{ pour tout } z \text{ et } F \neq G.$$

Si on rejette H_0 , on dit alors que le nombre de cases cochées pour un questionnaire «X» est stochastiquement plus grand que le nombre de cases cochées pour un questionnaire «Y». Pour Halifax, on utilise la contre-hypothèse $H_1: F^{(z)} \geq G^{(z)}$ pour tout z et $F \neq G$. Les résultats sont présentés dans le tableau 3. On peut remarquer que que la valeur de R_y dans la région de Québec pour chacune des questions est inférieure à 1. La raison est qu'à Québec, la plupart des gens n'ont coché qu'une case et que certains répondants n'ont pas répondu à l'une ou l'autre des questions.

Pour terminer, on a comparé les versions «X» et «Y» du deuxième essai modulaire quant au nombre de personnes s'identifiant comme d'origine ou d'identité française, italienne ou britannique dans certaines régions. Par britannique, on entend qu'au moins une des cases

Puisque $N_0^*/N_0 = N_1^*/N_1$, les tests sont équivalents. Afin de minimiser les calculs, il vaut mieux permuer les traitements si $N_1 < N_0$, et permuer les observations si $N_1' > N_0'$. Lors-que le nombre de permutations est grand, il a été suggéré par Dwass (1957) de tirer un échan-tilon de ces permutations et comparer la valeur observée de la statistique à l'ensemble des valeurs pour l'échantillon. Le fait de ne pas considérer toutes les permutations n'affecte pas le niveau du test mais seulement sa puissance. Si on échantillonne les permutations, la règle donnée plus haut tient toujours, non plus afin de minimiser les calculs, mais afin de minimi-ser la perte de puissance due à l'échantillonnage. Dwass montre, par exemple, que pour un test unilatéral au niveau 0.05, la perte de puissance avec un échantillon de 999 permutations est d'au plus 5.5%. Tel que mentionné dans Bradley (1968), les résultats d'une comparaison entre les tests aléatoires et les tests classiques quant à la puissance dépendront de la mesure dans laquelle les présupposés des tests classiques sont satisfais.

À cause de la façon dont sont construits les tests aléatoires, l'inférence ne porte que sur l'effi-cacité du traitement sur les unités de l'échantillon et non sur la population entière. D'un autre côté, les tests classiques sont basés sur un échantillon probabiliste tiré d'une population qui est rarement celle qui nous intéresse. Dans le cas qui nous occupe, par exemple, la population d'intérêt est la population canadienne le jour du recensement, le 4 juin 1991. Ainsi, pour les deux types de tests, des arguments non statistiques doivent être invoqués pour généraliser les inférences à la population d'intérêt.

4. UTILISATION DES TESTS ALÉATOIRES POUR L'ESSAI MODULAIRE 2

Comme il a été mentionné précédemment, le questionnaire du deuxième essai modulaire existe en deux versions: les versions «X» et «Y». La différence se situe au niveau des ques-tions sur l'origine et l'identité ethnique. En effet, dans la version «X», «CANADIEN» cons-titue une réponse possible (une case est réservée à cette fin) pour ces questions. La version «Y» n'offre pas ce choix de réponse pour ces mêmes questions, les personnes désirant répondre «CANADIEN» doivent inscrire cette réponse en toutes lettres après avoir coché la case «AUTRE(S)».

On désirait savoir si ces questions du questionnaire d'essai amènent plus ou moins de réponses multiples dans la version «X» que dans la version «Y». On dit qu'il y a une réponse multiple lorsque plus d'un choix de réponse est indiqué. On voulait aussi connaître l'influence du type de questionnaire sur la multiplicité (nombre de choix de réponse indiqués par le répondant) et sur certains choix de réponse (comme «FRANÇAIS») pour ces questions. Les types de ques-tionnaire constituent les traitements. Parce que l'échantillon de chaque région avait ses parti-cularités, les tests aléatoires ont été effectués séparément pour chacune des régions métropolitaines parmi lesquelles l'échantillon fut sélectionné.

Pour débiter, on génère aléatoirement un échantillon de 999 permutations de types de ques-tionnaire. Une permutation est générée de la manière suivante. Pour une région donnée, soient N_x et N_y le nombre de questionnaires «X» et «Y» respectivement. En utilisant l'algorithme de Bebbington (1975), on choisit parmi les $N_x + N_y$ ménages un échantillon aléatoire simple de N_x ménages auxquels on assigne le questionnaire de type «X». Ce processus est répété 999 fois. Par la suite, pour une question donnée, on calcule la proportion d'individus ayant donné une réponse multiple pour chacune des versions «X» et «Y» qu'on dénote P_x et P_y . Ensuite, pour chacune des 999 permutations de types de questionnaire ainsi que pour l'échan-tillon initial observé, on calcule la statistique $S = P_x - P_y$. On se retrouve donc avec 1000 valeurs de S que l'on classe en ordre croissant. Si, lors de cette classification, deux ou plusieurs statistiques sont égales, un nombre aléatoire entre 0 et 1 est généré et c'est ce nombre qui déter-minera l'ordre parmi les statistiques égales. La variable $RANGP_{x-y}$ représente le rang de la statistique S observée.

Valeurs de la statistique S pour chaque permutation des observations

Tableau 1

Traitement																Permutations																				
S	4/7				4/7				4/7				0				0				0				0				-4/15				-4/15			
	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1	1 2	1 2	0 1	0 1						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	2 3	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	1 2	2 3	1 2	1 2	0 1	1 2	0 1	1 2	0 1	2 3	0 1	2 3	0 1	2 3	1 2	2 3	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3	0 1	1 2	2 3						
X	2 3	1 2	1 2	2 3	2 3	1 2	2 3	2 3	1 2	2 3	1 2	2 3	1 2	2 3	0 1	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1	1 2	0 1								

Valeurs de la statistique S pour chaque permutation des traitements

Tableau 2

Observation		Permutations											
S	4/7	0				0				-4/15			
		Y	X	X	X	X	Y	Y	Y	X	X	X	
2 3	X	X	X	X	X	X	X	X	X	X	X	X	Y
1 2	X	X	X	X	X	X	X	X	X	X	X	X	X
0 1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	X

nombre d'unités qui reçoivent le traitement k ($k = 1, \dots, K$) et pour lesquelles on observe le résultat r_i ($i = 1, \dots, I$), $n_k = \sum_i n_{ki}$ le nombre d'unités recevant le traitement k , $n_i = \sum_k n_{ki}$ le nombre d'unités pour lesquelles on observe le résultat r_i et $n_{..} = \sum_k \sum_i n_{ki}$ le nombre total d'unités. Le nombre, N_i , de permutations des traitements est donné par

(1)
$$N_i = n_{..i} / \prod_k (n_{ki}).$$

De ces N_i permutations, il y en a N_i^* pour lesquelles il y a n_{ki} unités associées au traitement k et au résultat r_i ($k = 1, 2, \dots, K$; $i = 1, 2, \dots, I$), où

(2)
$$N_i^* = \prod_i \binom{n_{..i}}{n_{ki}} / \prod_k (n_{ki}).$$

D'autre part, il y a N_o permutations des observations, où

(3)
$$N_o = n_{..i} / \prod_i (n_{..i}).$$

De ces N_o permutations, il y en a N_o^* pour lesquelles on a n_{ki} unités associées au traitement k et au résultat r_i ($k = 1, 2, \dots, K$; $i = 1, 2, \dots, I$), où

(4)
$$N_o^* = \prod_k \binom{n_{k..}}{n_{ki}} / \prod_i (n_{ki}).$$

reformulée, ne pose pas de problèmes de compréhension auprès des répondants. On parle d'essais modulaires car ces enquêtes sont indépendantes les unes des autres et portent sur des parties distinctes du contenu du questionnaire du recensement.

Le premier essai, qui s'est tenu au mois de novembre 1987, avait comme objectif de mettre au point de nouvelles questions portant sur la couverture de la population, l'état matrimonial, la fécondité, le travail bénévolé ainsi que sur la nuptialité. Cette première enquête n'a fait l'objet d'aucun test, classique ou aléatoire.

Le deuxième essai, qui s'est tenu au mois de janvier 1988, visait principalement à mesurer la réaction des populations ethniques face à certaines questions portant sur la langue, l'origine ethnique, la religion, la citoyenneté ainsi que sur la mobilité. L'échantillon d'environ 3,500 ménages du deuxième essai modulaire provient d'un échantillonnage à deux degrés effectué à l'intérieur des régions métropolitaines de Halifax, Québec, Montréal, Toronto, Winnipeg et Vancouver. Pour des raisons de coût et pour faciliter la collecte des données, mais aussi parce qu'on voulait que l'échantillon contienne plusieurs personnes d'origines ethniques diverses, une méthode de sélection non probabiliste a été retenue. Le questionnaire du deuxième essai modulaire existe en deux versions dont les différences sont décrites à la section 4. Les ménages sélectionnés dans l'échantillon ont reçu aléatoirement un ou l'autre de ces questionnaires.

Pour nous permettre de tester statistiquement certaines hypothèses quant à ce deuxième essai modulaire, on a utilisé la méthode des tests aléatoires. Les tests aléatoires peuvent être utilisés pour comparer deux traitements appliqués aux unités d'échantillons qui ne sont pas nécessaires à la procédure utilisée pour un test aléatoire. On calcule la valeur d'une statistique pour les données observées, puis on calcule la valeur de la même statistique pour les autres permutations des données qui sont possibles avec le plan d'expérience utilisé. On rejette H_0 , si on juge que la valeur de la statistique pour les données observées est extrême par rapport aux valeurs obtenues sous H_0 pour l'ensemble des permutations.

3. TESTS ALÉATOIRES

Voici la procédure utilisée pour un test aléatoire. On calcule la valeur d'une statistique pour les données observées, puis on calcule la valeur de la même statistique pour les autres permutations des données qui sont possibles avec le plan d'expérience utilisé. On rejette H_0 , si on juge que la valeur de la statistique pour les données observées est extrême par rapport aux valeurs obtenues sous H_0 pour l'ensemble des permutations.

Par exemple, supposons qu'il y a quatre ménages: le premier ménage est formé de trois personnes, les ménages deux et trois comptent deux personnes chacun, et le quatrième ménage en compte une. Ces ménages peuvent avoir été sélectionnés arbitrairement, mais on choisit de façon aléatoire un ménage aux membres duquel on applique le traitement X , et on applique le traitement X aux membres des trois autres. Disons que le quatrième ménage est choisi pour le traitement X , que pour le premier ménage le traitement a réussi pour deux des trois personnes, tandis que pour les ménages deux et trois le traitement a réussi pour une des deux personnes, tandis que pour la personne du quatrième ménage le traitement n'a pas réussi. Notre hypothèse nulle est que les résultats sont indépendants du traitement employé. Pour mesurer l'impact du traitement X par rapport au traitement Y , on calcule la statistique S qui représente le nombre moyen de réussites pour le traitement X moins le nombre moyen de réussites pour le traitement Y . Dans notre cas $S = (2 + 1 + 1)/(3 + 2 + 2) - 0/1 = 4/7$. Pour savoir si cette valeur est significative, on présente dans le tableau 1 les valeurs de S obtenues en permettant les observations. Ici, une observation est constituée du nombre de personnes dans le ménage (après la barre verticale dans le tableau) et du nombre de réussites dans le ménage (devant la barre verticale dans le tableau). Si on fait un test à droite, on rejette H_0 si $\alpha \geq 3/12 = .25$, puisque trois des douze permutations donnent une valeur de S supérieure ou égale à $4/7$, la valeur observée.

Plutôt que de permuer les observations, on aurait pu permuer les traitements. Le tableau 2 montre ce qu'on obtient dans ce cas. Encore une fois, pour un test à droite, on rejette H_0 si $\alpha \geq 1/4 = .25$, puisque une seule des quatre permutations donne une valeur de S supérieure ou égale à $4/7$. Ce n'est pas une coïncidence si le résultat est le même. En effet, notons n_{ki} le

Un exemple d'utilisation de tests aléatoires pour les essais du questionnaire du recensement

YVES BÉLAND et ALAIN THÉBERGE¹

RÉSUMÉ

L'Essai modulaire 2, une enquête de Statistique Canada dont le but était d'aider à la mise au point du questionnaire du recensement de 1991 utilisait deux questionnaires distincts. L'échantillon de l'enquête était non probabiliste. Cet article décrit brièvement la méthodologie de l'enquête et comment les tests aléatoires ont été utilisés pour comparer les deux questionnaires.

MOTS CLÉS: Tests aléatoires; échantillonnage non probabiliste; plan d'expérience.

1. INTRODUCTION

On peut distinguer deux types de tests statistiques. Il y a ceux pour lesquels la décision est basée sur une comparaison de la valeur observée d'une statistique avec la distribution, sous l'hypothèse nulle, des valeurs de la même statistique pour l'ensemble des échantillons qui auraient pu être tirés. On les appellera les tests classiques. Pour faire un tel test, il faut connaître la probabilité d'obtenir un échantillon quelconque. Il doit donc y avoir eu échantillonnage probabiliste selon un plan de sondage connu. D'autre part, il y a les tests pour lesquels la décision est basée sur une comparaison de la valeur observée d'une statistique avec la distribution, sous l'hypothèse nulle, des valeurs de la même statistique pour l'ensemble des permutations possibles des données. On les appellera les tests aléatoires. Fisher (1935) en a fait l'utilisation pour comparer deux échantillons de graines, et Edgington (1987) traite de plusieurs aspects de ces tests. Les tests aléatoires requièrent d'abord l'existence de "traitements" afin de définir les permutations, ensuite il faut connaître la probabilité d'obtenir une permutation quelconque. On doit donc avoir choisi de façon aléatoire quelle unité allait être soumise à quel traitement, c'est-à-dire qu'il faut un plan d'expérience avec randomisation. Dans un organisme comme Statistique Canada, ce sont habituellement des tests classiques qui sont employés. Cet état de choses reflète d'une part, le fait que la plupart des enquêtes-échantillon effectuées à Statistique Canada utilisent un échantillonnage probabiliste, et d'autre part, l'absence de "traitements" dans ces enquêtes. On décrira dans cet article une enquête faisant exception à la règle, et comment on a appliqué les tests aléatoires.

Dans la section deux, on décrit brièvement la méthodologie employée pour les essais modulaires. La troisième section décrit à l'aide d'exemples simples la procédure utilisée pour un test aléatoire. L'application des tests aléatoires à l'Essai modulaire 2 est présentée dans la section quatre.

2. ESSAIS MODULAIRES

Dans le cadre de la planification du recensement de 1991, deux essais modulaires ont été effectués dans le but de valider les questions qui seront susceptibles d'être posées lors de ce recensement. Ces enquêtes visent à s'assurer que le libellé de chaque question nouvelle ou

¹ Yves Béland, Division des méthodes d'enquêtes sociales, Alain Théberge, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

donneurs et par conséquent, une bonne partie des classes d'imputation devraient se retrouver dans la catégorie "omnibus". Toutefois, ce genre de scénario n'est pas souhaitable en pratique. On peut simplifier le problème si l'on suppose que la répartition par âge est la même chez les hommes et chez les femmes pour la plupart des diagnostics. Des résultats de tests de signification permettent de croire qu'il ne s'agit pas là d'une hypothèse irréaliste. Dans l'exemple du SMH, nous avons décidé de grouper les diagnostics en fonction des valeurs estimées de μp et σp calculées à partir des données agrégées pour les deux sexes.

Notons aussi que nous ne nous sommes pas interrogés sur la pertinence de se servir de la moyenne et de l'écart type de la distribution selon l'âge pour attribuer des cotes numériques à chaque classe d'imputation. On pourrait aussi utiliser des percentiles ou d'autres paramètres de la distribution. Evidemment, les résultats de l'application de l'analyse typologique au problème du regroupement de classes dépendront du choix des cotes numériques.

Enfin, il est possible d'étendre la méthode proposée aux cas où $k \geq 1$ variables doivent être imputées ($k1$) et où $p \geq 2$ variables auxiliaires sont disponibles ($p2$).

REMERCIEMENTS

Cette étude a été présentée au congrès annuel de l'Association canadienne-française pour l'avancement des sciences en mai 1988. L'auteur tient à remercier Avi Singh qui, par ses précieux commentaires, a contribué à améliorer cet article. Il remercie également Cyril Nair de la Division de la santé ainsi que les membres de son équipe pour le soutien qu'ils lui ont offert, plus particulièrement en ce qui a trait à la production des fichiers informatiques requis à la réalisation de cette étude.

BIBLIOGRAPHIE

- ANDERBERG, M.R. (1973). *Cluster Analysis for Application*. New York: Academic Press.
- BALL, G.H., et HALL, D.J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics*, 12, 17-31.
- CORMACK, R.M. (1971) A review of classification. *Journal of the Royal Statistical Society*, série A, 134, 321-367.
- DREW, J.D., BÉLANGER, Y., et FOY, P. (1985). La stratification dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 11, 109-124.
- EVERITT, B.S. (1980). *Cluster Analysis*. Second Edition, London: Heineman Education Books Ltd.
- JUDKINS, D.R., et SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-284.
- HARTIGAN J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.
- MACQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings 5th Berkeley Symposium*, 1, 281-297.
- SAS INSTITUTE INC. (1985). *SAS User's Guide: Statistics*, Version 5.
- STATISTIQUE CANADA (1986). *Hospital Morbidity 1981-82, 1982-83*. Catalogue n° 82-206, Statistique Canada, Ottawa.
- WORLD HEALTH ORGANIZATION (1977). *International Classification of Diseases*. 1975 Revision, Volume 1, Genève.

Etape IV : Mise en application de FASTCLUS

En premier lieu, nous avons déterminé une manière de répartir les observations en 36 groupes (ce qui revenait à choisir 36 noyaux de grappe). On obtenait de meilleurs résultats en choisissant soi-même les noyaux de grappe initiaux qu'en laissant FASTCLUS les déterminer. Il convient de souligner que les résultats varieront selon l'ensemble initial de noyaux de grappe et l'ordre des données d'entrée puisque la méthode ne produit que des partitions localement optimales. Pour choisir les noyaux de grappe, nous avons réparti les diagnostics en neuf groupes de taille comparable selon m_d et en quatre groupes de taille comparable selon s_d . Nous avons pu ainsi obtenir 36 groupes homogènes de taille comparable, les valeurs moyennes des variables m_d et s_d dans chaque groupe servant de noyaux de grappe initiaux. Nous avons essayé plusieurs autres procédures et avons choisi celle qui donnait le coefficient R^2 le plus élevé.

En deuxième lieu, comme m_d et s_d reposaient sur un nombre d'observations variant beaucoup d'un diagnostic à l'autre, nous avons cru bon de procéder à une analyse typologique pondérée en nous servant comme poids du nombre d'observations disponibles pour chaque diagnostic. Notons que dans ce cas, FASTCLUS minimiserait la somme des carrés intra-grappe pondérée plutôt que la somme non pondérée.

3.4 Efficacité relative de la méthode proposée

Pour comparer la méthode de regroupement de classes d'imputation actuelle avec celle que nous proposons, nous pouvons nous servir du coefficient R^2 calculé pour l'ensemble des variables (dans le cas présent, ces variables seraient la moyenne et l'écart type). Le coefficient R^2 "combine" représente la proportion de la variance totale qui est expliquée par la somme des carrés inter-grappe "combine" (cette proportion devrait être aussi élevée que possible). Chaque somme "combine" des écarts est définie comme étant $(SSQ_m + SSQ_s)/2$, où SSQ_m et SSQ_s sont la somme des carrés pour la moyenne et pour l'écart type respectivement. Par la procédure FASTCLUS on obtient un coefficient R^2 de 0.993 pour m_d et de 0.929 pour s_d , un coefficient F^2 "combine" de 0.986. Avec la méthode de regroupement actuelle, on obtiendrait un coefficient R^2 de 0.735 pour m_d et de 0.466 pour s_d , pour un coefficient R^2 "combine" de 0.705. Ces résultats indiquent R^2 par conséquent que l'analyse typologique produit des groupes de diagnostics beaucoup plus homogènes par rapport à la variable à imputer que ne le fait la méthode actuelle.

4. CONCLUSIONS

En se servant de l'analyse typologique, nous avons présenté une méthode de regroupement des classes d'imputation d'une matrice d'imputation définie par le croisement de plusieurs variables auxiliaires. La méthode proposée a servi à l'imputation de l'âge pour le Système de morbidité hospitalière, où le diagnostic et le sexe étaient utilisées comme variables auxiliaires. Précisons que dans ce cas-ci, seule la variable "diagnostic" a été utilisée pour le regroupement des classes d'imputation originales. La variable "sexe" est néanmoins considérée dans le schéma d'imputation de manière à ce que l'enregistrement receveur et l'enregistrement donneur correspondent se rapportent à des personnes du même sexe. Dans une version généralisée de la méthode proposée, les deux variables pourraient servir au processus de regroupement. Si tel était le cas, il faudrait peut-être exiger que pour un diagnostic donné, les patients de sexe masculin et de sexe féminin se retrouvent dans une même rangée de la matrice d'imputation finale. Par ailleurs, on pourrait créer deux matrices d'imputation initiales serait nettement plus élevé, ce qui compliquerait davantage le problème de regroupement. Dans une telle situation, beaucoup de classes pourraient ne contenir qu'un petit nombre d'enregistrements

Etape I: Choix d'une méthode de groupement

Avant de choisir une méthode, mentionnons que notre objectif est essentiellement de répartir les diagnostics en groupes homogènes sans chercher à découvrir des grappes "naturelles" ou "réelles". C'est ce qu'on appelle dans la littérature statistique de la "dissection de données" (Everitt 1980). De plus, nous devons nous assurer que nous disposons d'un programme de typologie éprouvé qui emploie une méthode de groupement efficace. Le facteur déterminant dans le choix de la méthode a été le nombre d'observations contenues dans notre ensemble de données; c'est pourquoi nous avons choisi une technique disjointe plutôt que hiérarchique.

Compte tenu des remarques précédentes, nous avons opté pour la technique disjointe utilisée par la procédure FASTCLUS de SAS (1985) pour réaliser notre analyse. La procédure FASTCLUS exécute une analyse typologique suivant la distance euclidienne habituelle, calculée à partir d'un ensemble de variables quantitatives. Cette procédure comprend à la fois une méthode efficace pour déterminer les grappes initiales (celles-ci pouvant aussi être définies par l'utilisateur) et un algorithme d'itération standard permettant de minimiser la somme des carrés des distances par rapport aux moyennes de grappes. FASTCLUS s'inspire directement de l'algorithme principal de Hartigan (1975) et de l'algorithme à k -moyennes de MacQueen (1967). On choisit tout d'abord une série de noyaux de grappe, qui tiennent lieu d'estimations préliminaires pour les moyennes de grappe. On assigne ensuite chaque observation au noyau le plus près pour former des grappes temporaires. Chaque fois qu'une observation est assignée, le noyau de grappe est remplacé par la moyenne de la grappe temporaire (il s'agit là d'une opération facultative que nous avons choisie pour notre analyse). Après chaque traitement de l'ensemble de données, les observations sont assignées au noyau de grappe le plus près jusqu'à ce que les changements dans les valeurs des noyaux deviennent petits ou nuls (on a choisi qu'ils soient nuls dans notre cas). On forme les grappes finales en assignant chaque observation au noyau le plus près.

Etape II: Estimation des paramètres

Les données du SMH pour les années financières 1982-1983 et 1983-1984 ont servi à établir les estimations de m_d et de s_d pour chaque diagnostic d . Il s'agit là des estimations pondérées usuelles pour deux ans. Chaque diagnostic est représenté par deux variables, m_d et s_d . Il s'agit ici de trouver une répartition adéquate des diagnostics en fonction de considérations comme des valeurs aberrantes. Ces trois groupes correspondent aux trois premières rangées de la matrice d'imputation (les colonnes étant définies par la variable "sexe"). La dernière rangée de la matrice est réservée aux diagnostics pour lesquels il existe moins de dix observations dans les deux années en question et qui ne sont pas inclus dans l'un ou l'autre des trois groupes particuliers. La limite de dix observations a été fixée de façon arbitraire. Nous pouvons maintenant recourir à l'analyse typologique pour classer les diagnostics qui ne font pas partie des trois groupes particuliers et pour lesquels on a pu relever au moins dix observations dans les deux années étudiées.

Etape III: Détermination du nombre de grappes

La détermination du nombre de grappes a été faite en fonction de certaines contraintes opérationnelles puisque la matrice utilisée dans le programme d'imputation ne peut accepter plus de 40 rangées. Comme il y a déjà trois rangées de réserves aux diagnostics spécifiques et une aux diagnostics pour lesquels il existe moins de dix observations, nous pouvons encore utiliser 36 lignes sans que cela ne pose de problèmes. Selon une petite étude empirique ou l'on a calculé le coefficient R^2 pour divers nombres de grappes, la valeur de R^2 pour 36 grappes dépasse déjà 98%, ce qui semble indiquer que ce nombre soit raisonnable. Notons que pour aussi peu que 15 grappes, le R^2 atteint déjà 95%. Le coefficient R^2 est défini à la section 3.4.

nombre de groupes a été déterminé a posteriori de façon arbitraire. La principale faiblesse de cette méthode est qu'on ne s'est pas servi d'aucun critère statistique pour grouper les diagnostics, ce qui en fait une méthode plutôt laborieuse et assez subjective. Rappelons que les groupes de diagnostics ont été obtenus par une simple comparaison d'histogrammes. Une évaluation de la méthode d'imputation actuelle a permis de constater que les groupes de diagnostics n'étaient pas toujours homogènes par rapport à F_p et qu'ils devaient par conséquent être révisés.

3.2 Méthode proposée

La méthode proposée peut être décrite brièvement dans les termes suivants. Nous examinons ici le cas où une seule variable quantitative doit être imputée. L'imputation de plus d'une variable à la fois est examinée à la section 4. Désignons par y la variable à imputer et par F_i la distribution de cette variable à l'intérieur de la classe i . Notons que les classes sont formées par le croisement de variables auxiliaires ayant été préalablement divisées en catégories si nécessaires. La première étape consiste à déterminer un ensemble de paramètres qui puissent bien représenter F_i dans chaque classe, par exemple les trois ou quatre premiers moments de F_i ou les percentiles. La seconde étape consiste à estimer ces paramètres à l'aide des données fournies par les répondants. Enfin, en utilisant une méthode d'analyse typologique appropriée sur les paramètres estimés, on peut réduire le nombre de classes de façon à ce que les classes regroupées soient homogènes par rapport aux paramètres représentant F_i .

Nous allons maintenant justifier le choix de cette méthode dans le cas de l'imputation de l'âge pour le Système de morbidité hospitalière (SMH). Considérons tout d'abord d'autres méthodes de regroupement possibles. On pourrait par exemple adopter une méthode semblable à celle utilisée initialement pour les données de 1974, c'est-à-dire grouper les diagnostics en fonction des distributions F_p mais en se servant pour cela d'un critère statistique au lieu de comparer manuellement des histogrammes. Les données seraient classées d'après le diagnostic de référence, le sexe et l'âge (10 groupes d'âge disons). Deux diagnostics seraient réunis dans une même classe, si la proportion de cas rattachés à chacun des diagnostics dans chaque groupe d'âge, p_1, \dots, p_{10} , était semblable, cette similitude étant établie à l'aide d'un critère comme la distance euclidienne ou une mesure du chi carré. Notons que l'utilisation d'une mesure du chi-carré entraînerait de longs calculs puisqu'aucun programme d'analyse typologique courant n'utilise cette mesure de distance. Cela implique qu'il faudrait calculer la distance chi-carré pour toutes les paires de diagnostics possibles. Une autre méthode consisterait à appliquer tout d'abord une technique de réduction comme l'analyse en composantes principales pour réduire la dimension des groupes d'âge, puis à déterminer si deux diagnostics peuvent être apparentés en se fondant sur les scores obtenus sur chacune des composantes principales. Un inconvénient majeur à toutes ces méthodes est le nombre minimal d'observations requis pour obtenir une estimation fiable de la répartition par âge de chaque diagnostic.

Compte tenu des difficultés mentionnées ci-dessus, nous avons choisi d'utiliser les deux ou trois premiers moments pour décrire approximativement F_p . Nous avons considéré au départ trois moments: la moyenne m_p , l'écart type s_p et le coefficient d'asymétrie b_p . Toutefois, l'analyse en composantes principales nous a permis de constater que b_p était superflu. Il suffirait alors de grouper les diagnostics selon la moyenne d'échantillonnage m_p et l'écart type d'échantillonnage s_p . L'analyse typologique semble une approche statistique tout à fait appropriée à ce genre de problème. Un avantage évident que présente cette méthode par rapport à celles fondées sur la répartition par âge est qu'on a besoin de beaucoup moins d'observations pour estimer avec précision deux moments que pour estimer la proportion de cas pour plusieurs groupes d'âge. À la section suivante, nous voyons comment appliquer en détail cette méthode au problème d'imputation de l'âge.

3.3 Étapes de la mise en application de la méthode proposée pour les données du SMH

La mise en application de la méthode de groupement que nous proposons pour l'imputation de l'âge dans le SMH se fait en quatre étapes.

partition qui puisse optimiser un critère préalable. Les techniques disjointes se distinguent entre elles par la façon d'obtenir une partition initiale et par le critère de classification qu'elles visent à optimiser. En règle générale, on commence par choisir un ensemble de points appelé noyaux de grappe, qui sert d'estimation préliminaire pour les moyennes de grappes. Un certain nombre de méthodes ont été proposées pour le choix de ces points (Anderberg 1973). Une fois les noyaux de grappes choisis, on assigne chaque entité au noyau le plus près (à cette fin, on utilise habituellement la distance euclidienne). On peut réviser les valeurs estimées des moyennes de grappes chaque fois qu'une entité est assignée à un noyau (MacQueen 1967) ou uniquement après que toutes les entités aient été assignées (Ball et Hall 1967). Une fois que l'on a déterminé la partition initiale (ce qui équivaut à déterminer un ensemble de noyaux de grappe et à assigner chaque entité au noyau le plus près), on recherche les entités dont le transfert d'un groupe à un autre aurait pour effet d'améliorer le critère de classification. On répète cette opération jusqu'à ce qu'on ait atteint le point où il n'est plus possible d'améliorer le critère de classification par quelque transfert que ce soit. On obtient alors un optimum local. C'est ce qu'Anderberg (1973) appelle le "tri selon le centroïde le plus près". En règle générale, il n'est pas possible de savoir si on a obtenu un optimum global.

3. APPLICATION: CRÉATION DE CLASSES D'IMPUTATION POUR LE SMH

3.1 Contexte

Le Système de morbidité hospitalière (Statistique Canada 1987) consiste en un relevé des personnes hospitalisées qui, au cours de l'année à laquelle les données s'appliquent, ont reçu leur congé d'hôpitaux généraux et spécialisés au Canada, à l'exclusion du Yukon et des Territoires du Nord-Ouest. Chaque enregistrement contient au moins un code de diagnostic, l'âge et le sexe du patient, la durée de séjour, etc. Le diagnostic qui figure en premier sur l'enregistrement est appelé le diagnostic de référence et est celui sur lequel reposent les totalisations des publications. Il peut être considéré comme le principal motif d'hospitalisation et est codé suivant la 9^{ième} édition de la Classification internationale des maladies (Organisation mondiale de la santé 1977), qui contient plus de 5000 diagnostics.

À l'heure actuelle, l'imputation de l'âge dans le SMH se fait par méthode hot deck. On prédit l'âge d'un patient y à l'aide de deux variables auxiliaires, soit le diagnostic de référence d , qui est toujours indiqué sur l'enregistrement, et le sexe du patient s . Si le sexe n est pas indiqué, il faut l'imputer au préalable suivant le ratio hommes-femmes observée pour un diagnostic d'au d cours des années antérieures. En classant les patients en fonction de d et de s , on obtient une matrice d'imputation qui compte plus de 5000×2 classes d'imputation. Afin de réduire la dimension de cette matrice, nous avons groupé ou fusionné certains diagnostics à partir de la répartition par âge de chaque diagnostic. Soit F_d la répartition par âge dans la population des patients ayant le diagnostic de référence d . Alors, on groupera les diagnostics A et B si F_A est près de F_B . Pour ce faire, on peut se servir des estimations de F_d établies à l'aide des données disponibles. Il convient de souligner que nous n'avons pas utilisé la variable "sexe" pour définir les classes d'imputation (voir section 4 pour savoir comment elle pourrait être utilisée) même si elle est incluse dans le schéma d'imputation. En agissant ainsi, nous avons réduit de moitié le nombre de classes d'imputation contenues dans la matrice.

Afin de permettre une appréciation plus juste de la méthode que nous proposons pour grouper les classes d'imputation, nous allons tout d'abord décrire la méthode utilisée actuellement et en exposer les limites. Le regroupement s'est fait en comparant manuellement (à l'aide d'histogrammes) la forme des distributions par âge empiriques, F_d , relatives à tous les codes de diagnostic qui correspondent aux données du SMH de 1974. Trente-six groupes ont été ainsi formés et un 37^{ième} a été créé pour les diagnostics ayant 200 observations disponibles. Le

fait en fonction de deux variables auxiliaires: le sexe et le diagnostic. On compte plus de 5000 classes d'imputation pour chaque sexe. Nous proposons de réduire le nombre de classes d'imputation au moyen de l'analyse typologique en groupant certains niveaux de la variable "diagnostic" de manière à obtenir 40 groupes de diagnostics. À la section 2, nous faisons un survol des méthodes d'analyse typologique les plus courantes. À la section suivante, nous appliquons l'analyse typologique au groupement de classes d'imputation en prenant l'exemple de l'imputation de l'âge dans le SMH; nous comparons de plus l'efficacité de la méthode proposée à celle de la méthode utilisée actuellement. Les deux méthodes sont de type "hot deck" sauf que la méthode que nous proposons redéfinit les classes d'imputation au moyen de l'analyse typologique. Enfin à la section 4, nous présentons quelques remarques globales et proposons une généralisation de la méthode.

2. APERÇU DES MÉTHODES D'ANALYSE TYPOLOGIQUE

Dans cette section, nous examinons le problème consistant à classer un nombre donné d'entités définies par un certain nombre de variables quantitatives de façon à ce que les entités d'un même groupe ou d'une même grappe soient semblables entre elles et différentes des entités des autres groupes. Everitt (1980) présente un bon sommaire des méthodes d'analyse typologique basé principalement sur l'article de Cormack (1971). La plupart des méthodes de d'analyse typologique appartiennent à l'une ou l'autre de deux catégories: techniques hiérarchiques et techniques disjointes, ces dernières étant aussi appelées techniques d'optimisation. Nous décrivons plus bas ces deux catégories. Parmi d'autres méthodes, mentionnons les techniques de densité, où l'on forme des grappes en recherchant les régions qui renferment de fortes concentrations d'entités. En effet, si les entités sont définies comme les points d'un espace métrique, il devrait y avoir des portions de cet espace où la densité de points est très élevée et d'autres où elle est plus faible. Notons aussi les techniques d'agglomération, où il peut y avoir chevauchement des grappes. À titre d'exemple, dans certaines disciplines comme la linguistique, les mots peuvent avoir plusieurs significations. Ainsi si l'on regroupe les mots suivant leur signification, ceux-ci devraient appartenir à plusieurs catégories se chevauchant entre elles.

Les techniques hiérarchiques se divisent en deux groupes: techniques de groupement par fusion et techniques de groupement par division. Dans le premier cas, chaque entité constitue au départ une grappe. Par la suite, les deux grappes les plus près l'une de l'autre sont fondues en une seule; on procède ainsi jusqu'à ce qu'il ne reste plus qu'une seule grappe contenant toutes les observations. Dans le deuxième cas, toutes les entités sont contenues au départ dans une seule grappe. Par la suite, cette grappe est décomposée graduellement en plus petites grappes jusqu'à ce que chaque entité constitue en soi une grappe. Les techniques hiérarchiques se distinguent les unes des autres suivant la définition de la distance entre les observations ou des groupes d'observations. De plus, grâce à ces techniques, on peut obtenir en une seule phase d'exécution des résultats pour n'importe quel nombre de grappes désiré en interrompant le processus de fusion ou de division au niveau hiérarchique voulu. Évidemment, les techniques hiérarchiques ne peuvent servir que pour de petits ensembles de données puisqu'il existe $n(n-1)/2$ façons de grouper deux entités parmi n et $2^{n-1} - 1$ façons de diviser un groupe de n entités en deux autres groupes.

Contrairement aux techniques hiérarchiques, où les observations appartiennent à une série de grappes suivant le niveau de la hiérarchie, les techniques disjointes répartissent les observations en un certain nombre de grappes (généralement préalable) de telle manière que chaque observation appartient à une et une seule grappe. Elles se distinguent aussi des techniques hiérarchiques en ce qu'elles permettent une redistribution des observations, de sorte qu'il est possible de corriger ultérieurement une répartition initiale qui laisse à désirer. Les techniques disjointes conviennent nettement mieux que les techniques hiérarchiques pour de grands ensembles de données. On les appelle aussi techniques d'optimisation parce qu'elles recherchent une

Analyse typologique appliquée au regroupement de classes d'imputation

E.R. LANGLET¹

RÉSUMÉ

Dans cet article, nous nous intéressons au problème du regroupement de classes d'imputation définies par un grand nombre de combinaisons de variables auxiliaires. En nous fondant sur les principes de l'analyse typologique, nous proposons une solution qui vise à réduire le nombre de niveaux des variables auxiliaires de manière à obtenir un nombre de classes d'imputation plus raisonnable. À titre d'exemple, nous examinons le processus d'imputation de l'âge dans le Système de morbidité hospitalière, où les variables auxiliaires sont le sexe et le diagnostic.

MOTS CLÉS: Non-réponse partielle; variables auxiliaires; matrice d'imputation; enregistrements donneurs; techniques disjointes; techniques hiérarchiques; noyaux de grappe.

1. ÉNONCÉ DU PROBLÈME

La non-réponse partielle survient dans les enquêtes lorsque l'on ne recueille pas toutes les données voulues sur une unité d'échantillonnage ou qu'on doit éliminer certaines données parce qu'elles ne respectent pas les conditions de contrôle. Dans de nombreuses enquêtes, on résout cette difficulté en recourant à l'imputation aléatoire à l'intérieur de classes, une version courante de la méthode hot deck. Pour réaliser ce genre d'imputation, on choisit aléatoirement un répondant dans une classe d'imputation définie par une ou plusieurs variables auxiliaires et on impute au non-répondant la valeur associée à ce répondant.

Nous pouvons exposer le problème de la façon suivante. Les répondants sont classés selon certaines variables auxiliaires; ces classes ou groupes forment une matrice d'imputation multidimensionnelle où le nombre de cellules définies par la combinaison des variables auxiliaires est égal au nombre de classes d'imputation. Si ce nombre est très élevé, plusieurs classes pourraient ne pas contenir d'enregistrements donneurs ou en contenir très peu. En outre, la manipulation d'une telle matrice peut devenir très laborieuse numériquement. Une façon de contourner cette difficulté est de réduire le nombre de cellules de la matrice soit en groupant des cellules, des rangées ou des colonnes de cette matrice ou des combinaisons de celles-ci de manière à obtenir des groupes homogènes par rapport aux variables à imputer. Pour réaliser cela, nous proposons de recourir à l'analyse typologique. À cette fin, nous pouvons attribuer une cote numérique à chaque classe d'imputation en nous servant de la valeur des variables d'intérêt tirées des enregistrements donneurs (ou répondants) pour chaque classe. Dans cet article, nous quantifions les classes d'imputation à l'aide de mesures fondées sur la distribution empirique des variables à imputer obtenue à partir des données des répondants. Nous pouvons ensuite utiliser l'analyse typologique pour grouper les cellules de la matrice en fonction des cotes numé-riques attribuées. Nous verrons que l'analyse typologique est appropriée dans de telles circonstances. D'autres applications similaires de l'analyse typologique à la stratification d'unités primaires d'échantillonnage sont présentées dans Drew, Bélanger et Foy (1985), Judkins et Singh (1981) ainsi que les ouvrages qui y sont cités.

Le problème auquel nous nous intéressons ici trouve son origine dans l'imputation de l'âge pour le Système de morbidité hospitalière (SMH). Dans ce système, l'imputation de l'âge se

¹ E.R. Langlet, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, K1A 0T6.

- BRADBURN, N.M., et SUDMAN, S. (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- CANNELL, C.F., et FOWLER, F.J. (1963). Comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly*, 27, 250-264.
- DELAMETER, J., et MacCORQUODALE, P. (1975). The effects of interview schedule on reported sexual behavior. *Sociological Methods and Research*, 4, 215-236.
- ELLIS, A. (1947). Questionnaire versus interview methods in the study of human love relationships. *American Sociological Review*, 12, 541-553.
- GIBSON, F.W., et HAWKINS, B.W. (1968). Interview versus questionnaires. *American Behavioral Scientist*, 12, 9-11.
- HERZOG, A.R., RODGERS, W., et KULKAR, R.A. (1983). Interviewing older adults: a comparison of telephone and face-to-face modalities. *Public Opinion Quarterly*, 47, 405-418.
- HETHERINGTON, R.W., DICKINSON, J., CIPRYWNYK, D., et HAY, D.A. (1978). Drinking behavior among Saskatchewan adolescents. *Canadian Journal of Public Health*, 69, 315-324.
- HETHERINGTON, R.W., DICKINSON, J., CIPRYWNYK, D., et HAY, D.A. (1979). Attitudes and knowledge about alcohol among Saskatchewan adolescents. *Canadian Journal of Public Health*, 70, 247-259.
- HUBBARD, R.L., ECKERMAN, W.C., et RACHAL, J.V. (1976). Methods of validating self-reports of drug use: a critical review. *Proceeding of the Social Statistics Section, American Statistical Association, Part I*, 406-409.
- KNUDSEN, D., HALLOWELL, D., et IRISH, D.P. (1967). Response differences to questions on sexual standards: an interview-questionnaire comparison. *Public Opinion Quarterly*, 21, 290-297.
- KROHN, M., WALDO, G.P., et CHIRICOS, T.G. (1974). Self-reported delinquency: a comparison of structured interviews and self-administered checklists. *The Journal of Criminal Law and Criminology*, 65, 545-553.
- LOCANDER, W., SUDMAN, S., et BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.
- MCDONAGH, E.C., et ROSENBLUM, A.L. (1965). A comparison of mailed questionnaires and subsequent structured interviews. *Public Opinion Quarterly*, 29, 131-136.
- METZNER, H., et MANN, F. (1952). A limited comparison of two methods of data collection: the fixed alternative questionnaire and the open-ended interview. *American Sociological Review*, 17, 486-491.
- NEWTON, R.R., PRENSKY, D., et SHUESSLER, K. (1982). Form effect in the measurement of feeling states. *Social Science Research*, 11, 301-317.
- SCHUMAN, H. (1980). Review of improving interview method and questionnaire design. *Social Forces*, 59, 325-326.
- SELLTZ, C., JAHODA, M., DEUTSCH, M., et COOK, S.W. (1959). *Research Methods in Social Relations* (Revised). New York: Holt, Rinehart and Winston.
- SIEMIATYCKI, J. (1979) A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health*, 69, 238-245.
- SMITH, H.W. (1975). *Strategies of Social Research*. Englewood Cliffs: Prentice Hall.
- SYKES, W.M., et COLLINS, M. (1987). Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni. *Techniques d'enquêtes*, 13, 19-33.
- WHITEHEAD, P.C., et SMART, R.G. (1972). Validity and reliability of self-reported drug use. *Canadian Journal of Criminology and Corrections*, 14, 83-89.
- WISEMAN, F. (1972). Methodological bias in public opinion polls. *Public Opinion Quarterly*, 36, 105-108.

recevoir des notes scolaires significativement plus élevées, avaient de plus grandes aspirations pour leurs études futures et ont révélé une image de soi plus positive pour 4 des 7 éléments d'estime de soi et pour l'indice composite d'estime de soi. Contrairement à l'hypothèse selon laquelle les répondants à l'interview auraient dû vouloir donner d'eux une image plus favorable, ces résultats tendent à montrer que les étudiants interviewés se sont montrés plus modestes dans la déclaration des notes obtenues, de leurs aspirations scolaires et de leur perception d'eux-mêmes. Cependant, les étudiants qui ont rempli le questionnaire et ont plus volontiers déclaré leurs comportements liés à l'alcool en raison du caractère plus anonyme du questionnaire et de la plus grande liberté qu'il semblait offrir, se sont peut-être aussi sentis plus libres de réhausser leur image personnelle sur des sujets comme les notes obtenues, les aspirations scolaires et leur perception de soi.

Toutefois, la présence d'un biais de réponse entre les distributions tirées des données des interviews et des questionnaires est évidente uniquement au sens statistique du terme. Les écarts statistiquement significatifs entre les valeurs moyennes pour les questions examinées se situaient entre 0.05 et un maximum de 0.48 dans le cas de l'indice composite d'estime de soi. Vu la présence possible d'autres erreurs de mesure, les différences entre les réponses obtenues par la méthode de l'interview et celle du questionnaire dans la présente étude n'ont pas une ampleur suffisante pour qu'on puisse les considérer comme des signes de l'existence d'un biais de réponse important d'un point de vue pratique.

Comme on ne disposait pas de renseignements fiables sur les habitudes réelles de consommation des étudiants et des parents, sur les notes scolaires et d'autres réponses examinées, il n'a pas été possible de procéder à une évaluation de l'exactitude relative des réponses à l'interview et au questionnaire. Par conséquent, il est impossible d'indiquer la supériorité relative soit du questionnaire rempli par les répondants, soit de l'interview sur place, pour les réponses qui ont retenu notre intérêt. Les deux types de réponses peuvent être sujets à un biais de sous-déclaration ou de surdéclaration de direction et d'ampleur indéterminées.

Les résultats exposés ici corroborent en général ceux de Bradburn et Sudman (1979), qui indiquent qu'il ne semble exister aucune relation systématique entre la méthode d'obtention des réponses et la surdéclaration de comportements socialement désirables ou la sous-déclaration de comportements et d'attitudes socialement indésirables. En conséquence, Bradburn et Sudman (1979) et Locander et coll. (1976) font valoir qu'aucune méthode de collecte des données n'est clairement supérieure pour tous les types de questions à composante "menaçante" ou autres aspects intéressant les spécialistes des enquêtes.

REMERCIEMENTS

L'auteur exprime ses remerciements à S. Parvez Wakil pour son examen critique de la version originale de cet article, ainsi qu'aux lecteurs anonymes et à M. P. Singh pour leurs commentaires très utiles.

L'étude initiale a reçu l'appui de la Direction de l'usage non médical des drogues de Santé et Bien-être social Canada (subvention n° 1213-7-10) et un soutien additionnel de l'Unité de recherche appliquée, Division de la recherche psychiatrique, ministère de la Santé de Saskatchewan. Les auteurs de l'étude initiale sont également remerciés pour la générosité avec laquelle ils ont permis d'utiliser leurs données pour les fins du présent article.

BIBLIOGRAPHIE

- ALWIN, D.F. (1977). Making errors in surveys: an overview. *Sociological Methods and Research*, 6, 131-151.
- BLAIR, E., SUDMAN, S., BRADBURN, N.M., et STOCKING, C. (1977). How to ask questions about drinking and sex: response effects in measuring consumer behavior. *Journal of Marketing Research*, 14, 316-321.

Tableau 2

Moyennes, écarts types et probabilités "t" pour certaines questions posées aux répondants à l'interview et au questionnaire

Variable ¹	Interview (n = 752)		Questionnaire (n = 750)	
	X	S	X	S

Consommation d'alcool et comportements connexes

Niveau de consommation	2.31	2.92	2.76	3.05
Âge au premier verre ^a	3.93	1.32	3.64	1.39
Le père boit	1.82	0.62	1.90	0.58
La mère boit	1.70	0.50	1.75	0.51
Les amis boivent	1.92	0.57	1.94	0.56

Variables scolaires

Notes obtenues	4.37	1.49	4.58	1.46
Plans scolaires	3.02	1.24	3.25	1.24

Variables religieuses

Importance de la religion à la maison	3.37	1.16	3.15	1.22
Importance de la religion pour l'étudiant	3.22	1.12	3.13	1.18

Indices d'estime de soi

Élément 1	2.98	0.60	3.12	0.60
Élément 2	2.96	0.49	3.08	0.54
Élément 3	3.14	0.55	3.27	0.61
Élément 4	2.98	0.51	2.05	0.57
Élément 5	3.10	0.63	3.01	0.75
Élément 6	2.93	0.56	2.97	0.59
Élément 7	3.07	0.54	3.12	0.60
Composite	21.17	2.39	21.65	2.85

a - Valeur moyenne calculée sur des données groupées.

¹ Codes des variables: Niveau de consommation: indice composite de la fréquence et du volume d'alcool consommé; 0 = ne consomme pas à 9 = consomme souvent et en grandes quantités.
Âge au premier verre: 1 = 6 ans ou moins; 2 = 7-8 ans; 3 = 9-10 ans; 4 = 11-12 ans; 5 = 13-14 ans; 6 = 15-16 ans et 7 = ≥ 17 ans et plus.
Le père, la mère ou les amis boivent: 1 = jamais; 2 = parfois; 3 = beaucoup.
Notes obtenues: 1 = surtout D et F; 2 = surtout C et D; 3 = surtout C; 4 = surtout B et C; 5 = surtout B; 6 = surtout A et B; 7 = surtout A.
Plans scolaires: 1 = ne finira pas la 12^e année; 2 = s'arrêtera après la 12^e année; 3 = recevra une formation technique; 4 = ira à l'université; 5 = fréquentera une école d'études supérieures ou professionnelle.
Éléments d'estime de soi et indice: 1 = fortement en désaccord; 2 = en désaccord; 3 = d'accord; 4 = entièrement d'accord. L'indice additif pour les 7 éléments allait de 7 à 28.

L'interview n'étaient pas plus susceptibles de déclarer que la religion était importante pour eux que les répondants au questionnaire. De même, en ce qui concerne les habitudes de consommation des amis ou des camarades de classe, les deux groupes de répondants étaient aussi susceptibles l'un que l'autre d'en faire état.
Les profils des réponses fournies à d'autres questions comportant des aspects quelque peu différents liés aux expériences personnelles et à l'image de soi n'ont pas, en général, permis de conclure en la présence d'un effet de désirabilité sociale, comme c'était le cas pour la consommation d'alcool. Comme l'indique le tableau 2, les répondants au questionnaire ont déclaré

Une comparaison des réponses moyennes ou des distributions de fréquences pour les répondants à l'interview et au questionnaire, dans le cas d'un certain nombre de questions à incidence non normative ou illégale, a corroboré de façon générale les recherches antérieures sur des sujets semblables. Les questions les plus préoccupantes sont celles ayant trait à la consommation d'alcool, dans lesquelles une forte composante de menace ou de déviance est perçue par les répondants visés, dont la majorité (99,8%) n'avaient pas l'âge légal pour consommer de l'alcool au moment de l'étude.

Les distributions de fréquences du tableau 1 montrent qu'un pourcentage significativement supérieur des répondants ont déclaré avoir déjà plus que goûté ou pris une petite gorgée d'une boisson alcoolique. Des écarts statistiquement significatifs du même ordre ont été observés entre les répondants à l'interview et les répondants au questionnaire au sujet de l'usage du tabac.

Dans le cas des répondants ayant déclaré avoir consommé un verre d'alcool, les niveaux moyens de consommation et l'âge moyen à la consommation du premier verre indiqués au tableau 2 semblent également indiquer que les répondants au questionnaire sont plus enclins à faire état de comportements déviants que les élèves du même âge ayant répondu à l'interview. Les niveaux moyens de consommation significativement plus élevés des répondants au questionnaire reflète la déclaration par ces derniers de quantités et de fréquences plus élevées de consommation d'alcool. L'âge moyen à la première consommation significativement plus élevée dans le cas des répondants à l'interview indique qu'ils ont déclaré avoir pris leur première consommation réelle à un âge plus avancé que les répondants au questionnaire.

Des différences significatives entre les répondants à l'interview et les répondants au questionnaire ont également été observées dans le cas de la consommation des parents et de l'importance de la religion à la maison. Les valeurs moyennes, pour ces trois questions, indiquent que les répondants au questionnaire ont déclaré des niveaux de consommation plus élevés pour leurs parents que ceux qui ont été interviewés, et que la religion était perçue comme moins importante à la maison par les répondants au questionnaire. Bien que ces questions ne comportaient pas, en tant que telles, le même caractère personnel ou le même aspect menaçant pour le répondant, les écarts ont été considérés comme une indication du fait que les étudiants interviewés tentaient de donner de leur vie familiale une image plus favorable et plus socialement acceptable. Toutefois, l'importance plus grande de la religion à la maison déclarée par les répondants à l'interview ne s'est pas répercutée sur leurs réponses quant à l'importance qu'ils accordent eux-mêmes à la religion. L'équivalence statistique des valeurs moyennes en ce qui touche l'importance de la religion pour les répondants eux-mêmes indique que les répondants à

2.1 Distribution des variables

Le groupe ethnique étant mis à part, on a donc pu fonder l'analyse subséquente sur l'hypothèse que les répondants à l'interview et au questionnaire étaient équivalents à l'égard de plusieurs variables qui auraient pu rendre confuse la comparaison des réponses obtenues selon les deux méthodes.

Variable	Interview (n = 752)	Questionnaire (n = 750)	Probabilité Z bilatérale
Déjà bu	62.63	73.73	.000
Déjà fumé la cigarette	29.78	37.60	.001

Distribution de fréquences et probabilités Z pour certaines questions posées aux répondants à l'interview et au questionnaire

Tableau 1

Sykes et Collins (1987). D'autres chercheurs ont observé qu'on était susceptible d'obtenir davantage de réponses naturelles, révélatrices et informatives par des questionnaires et des entretiens téléphoniques que par des entretiens personnels, en ce qui a trait à des attitudes ou à des comportements déviants, délicats ou embarrassants. (Cannell et Fowler 1963; Ellis 1947; Hubbard et coll. 1976; Knudsen et coll. 1967; Siemiatycki 1979; Whitehead et Smart 1972 et Wiseman 1972).

Les conclusions de ces dernières études découlent généralement de l'hypothèse non vérifiée qu'une plus grande déclaration de renseignements sur des aspects déviants, menaçants ou embarrassants correspondait à des résultats plus exacts (Blair et coll. 1977). Cet argument a également été invoqué par Schuman (1980), qui a affirmé que fréquemment aucune information de validation externe n'était obtenue, mais que les chercheurs "assumaient que plus de tels comportements étaient déclarés, plus les déclarations étaient exactes - hypothèse plausible mais non irrefutable pour la plupart des sujets traités".

Le présent article s'intéresse à une comparaison plus approfondie de la relation entre les entretiens sur place et les questionnaires remplis par les répondants et les réponses obtenues auprès d'une population d'adolescents sur un aspect "menaçant" ou déviant, soit la consommation d'alcool. Les résultats dont il est fait état sont fondés sur une analyse secondaire de données provenant d'une étude sur les attitudes et les comportements liés à l'alcool, menée auprès d'un échantillon d'adolescents dans une province de l'Ouest canadien en 1977-1978 (Hetherington et coll. 1978 et 1979).

Dans cette étude, tant la méthode de l'entrevue sur place que celle du questionnaire rempli par les répondants ont été utilisées, ce qui fournit une occasion unique de comparer les effets possibles du mode de collecte des données sur les données résultantes. Ce genre de comparaison, qui présente un intérêt pour les spécialistes des enquêtes, n'est pas possible dans la majorité des enquêtes, qui se fondent sur une méthode unique de collecte de l'information.

Un échantillon aléatoire stratifié de 1502 étudiants des niveaux de la 6^e à la 12^e année a été prélevé dans trois régions scolaires de la province concernée. L'échantillon total d'étudiants a été réparti au hasard, par niveau, entre la méthode du questionnaire rempli par les répondants et celle de l'entrevue sur place. Ainsi, environ la moitié des étudiants de chaque niveau (de la 6^e à la 12^e année) ont été soumis à chacune des méthodes. Le nombre d'étudiants affectés à l'entrevue sur place était de 752, tandis que celui des étudiants devant remplir le questionnaire était de 750.

Les étudiants remplissant le questionnaire l'ont fait en groupe sous la surveillance d'un chercheur qualifié, dans une salle réservée à cette fin dans chaque école. Les entretiens ont été menés par quinze interviewers ayant reçu une formation spéciale pour les fins de l'étude. Le contenu de l'enquête, qui comprenait 75 questions, était identique dans l'entrevue et dans le questionnaire. La majorité des questions étaient des questions fermées et il fallait en moyenne 20 minutes pour répondre à l'enquête selon l'une ou l'autre méthode.

2. RÉSULTATS ET ANALYSE

Une comparaison entre les répondants affectés à l'entrevue personnelle et ceux ayant rempli le questionnaire a été effectuée à l'égard d'un certain nombre de caractéristiques personnelles et familiales, afin de déterminer si les deux groupes présentaient des différences autres que la méthode par laquelle ils avaient fourni les données. Les résultats ont révélé que les deux groupes ne présentaient pas une différence plus grande que celle attribuable au hasard pour des variables comme le sexe, l'âge, l'année scolaire, le niveau de scolarité et la profession des parents, et l'appartenance religieuse. Une différence statistiquement significative a été observée dans le cas de la variable groupe ethnique, les répondants à l'entrevue ayant déclaré en plus forte proportion une origine canadienne.

Le choix de la méthode est-il important pour les sujets d'enquête délicats?

DAVID A. HAY¹

RÉSUMÉ

Les effets de l'utilisation d'un questionnaire à remplir soi-même ou d'une méthode d'interviews sur place sur les réponses d'un échantillon d'adolescents au sujet de leur consommation d'alcool et de leurs comportements connexes sont examinés. Les résultats corroborent en général les études précédentes sur la relation qui existe entre la méthode de collecte des données et la distribution des réponses présentant un contenu délicat ou non normatif. Bien que significatives sur le plan statistique, un bon nombre des différences ne sont pas assez grandes pour être considérées comme importantes sur le plan pratique.

MOTS CLÉS: Collecte des données; interview sur place; questionnaire rempli par le répondant; erreurs de réponse; consommation d'alcool.

1. INTRODUCTION

Procéder par questionnaire ou par interviews, voilà l'alternative qui se pose aux chercheurs au moment de concevoir et d'effectuer des enquêtes-échantillons portant sur des sujets de nature délicate. Le choix entre l'utilisation d'interviews personnelles ou par téléphone et le recours à une variante des questionnaires remplis par les répondants eux-mêmes, ou encore l'utilisation des deux à la fois, est une décision cruciale que les concepteurs d'une enquête doivent prendre dans le but d'optimiser la qualité des données obtenues.

Outre les problèmes généraux de fiabilité et de validité liés à la déclaration, par les intéressés eux-mêmes, d'attitudes, de comportements et d'autres éléments intéressant les spécialistes des enquêtes, il y a la question sous-jacente des capacités relatives de l'interview et du questionnaire rempli par le répondant de réduire au minimum ou de diminuer les biais ou erreurs non dus à l'échantillonnage. En d'autres termes, l'utilisation de modes différents de collecte des données produira-t-elle des résultats différents (Smith 1975)?

Dès 1959, Sellitz et coll. (1959) affirmaient que la plupart des questionnaires et des interviews étaient utilisés sans qu'on dispose d'informations sur leurs mérites relatifs. Cette position a été réaffirmée plus récemment par Knudsen et coll. (1967), Alwin (1977) et Newton et coll. (1982), qui prétendent que le choix du mode d'enquête est fondé sur la commodité, les coûts et d'autres considérations pratiques, plutôt que sur la correspondance entre la méthode et le sujet et les effets possibles sur les réponses. Selon Newton et coll., la planification des enquêtes devrait s'inspirer des données fiables connues au sujet de la relation entre les méthodes d'exécution et les profils de réponses, plutôt qu'être basée uniquement sur les coûts comparatifs, la motivation des répondants et d'autres aspects analogues.

Certaines études dans lesquelles on a fait la comparaison entre les interviews sur place et des formes plus anonymes comme les questionnaires remplis par les répondants ou les interviews téléphoniques ont révélé des différences minimes ou non significatives du point de vue statistique dans les réponses données à toute une gamme de questions, notamment sur des sujets de nature personnelle ou délicate (DeLameter et MacCorquodale 1975; Gibson et Hawkins 1968; Krohn et coll., 1974; McDonagh et Rosenblum 1965; Metzner et Mann 1952 et Newton et coll. 1982

¹ David A. Hay, Professeur agrégé, Université de la Saskatchewan, Saskatoon, Saskatchewan, S7N 0W0.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur en chef de la revue et les arbitres qui, par leurs précieux commentaires, ont contribué à l'amélioration de cet article.

BIBLIOGRAPHIE

- BETHEL, J. (1986). An optimum allocation algorithm for multivariate surveys. Rapport technique du United States Department of Agriculture, Statistical Reporting Service, Statistical Research Division, numéro SF et SRB-89.
- GERMAIN, M.-F., DOLSON, D., et MARANDA, F. (1989). Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988. Document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles, Statistique Canada.
- HARTIGAN, J.A. (1975). *Clustering Algorithms*. New York: John Wiley and Sons.
- INGRAM, S., et DAVIDSON, G. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 220-225.
- LYNCH, J. (1988). Cas spéciaux d'estimation dans l'enquête nationale des fermes de 1988. Document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles, Statistique Canada.
- MacQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.
- MASSART, D.L., et KAUFMAN, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley and Sons.
- SAS INSTITUTE INC. (1985). *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS institute.
- STATISTIQUE CANADA (1987). Dictionnaire du Recensement de 1986. Statistique Canada, n° 99-101F au catalogue.
- WARD, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

en 1988. Par contre, certaines de ces estimations sont beaucoup moins précises, et 7 % de l'ensemble sont même deux fois moins précises. Ces estimations proviennent du Québec et de l'Ontario où on collecte les dépenses d'exploitation seulement chez les fermes incorporées. Or, dans ces provinces, la forme juridique d'une ferme est difficile à identifier, autant lors du recensement que lors de l'enquête. Malgré la réduction de l'échantillon effectif en raison de la non-réponse totale et des coupures au moment du développement du plan de sondage, on conclut que l'enquête de 1988 a fourni en général des estimations plus précises pour chaque catégorie de variables.

9.2 Précision atteinte contre précision espérée

On s'attend à ce que la précision atteinte soit moins bonne que la précision espérée, et ce, pour deux raisons. Premièrement, l'ajustement des facteurs de pondération pour compenser la non-réponse totale entraîne une augmentation de la variance. Deuxièmement, les données qui ont servi à créer la base de sondage proviennent du recensement agricole de 1986. D'une part, ces données sont sujettes à des erreurs, et, d'autre part, la base de sondage se détériore, à la suite des changements dans l'activité agricole. La précision est comparée en employant le coefficient de variation des estimations du niveau provincial provenant de la base de liste L1 seulement. Ces estimations sont celles de plusieurs variables clés dont le CV espéré n'excédait pas 20 %.

La comparaison de la précision de 288 estimations est présentée à l'aide de graphiques de la figure 3. Sur ces graphiques, chaque carré représente, pour une estimation, le CV espéré, en abscisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des variables clés situées à l'intérieur de chaque zone délimitée par les droites $Y = X$, $Y = 2X$ et $Y = 3X$.

Pour les cultures et le bétail, environ 90 % des estimations ont une précision acceptable, compte tenu du taux de non-réponse, puisque la plupart des variables clés se situent plus près de la droite $Y = X$ que de la droite $Y = 2X$. Pour les dépenses, on distingue deux tendances. D'abord, fait remarquable, le CV atteint est inférieur dans 28 % des cas à celui qui était espéré; ces cas proviennent en grande majorité de la région de la CCB. Par contre, 31 % de l'ensemble présente une précision plus que deux fois plus faible qu'espérée. Ces cas proviennent du Québec et de l'Ontario pour les raisons mentionnées à la section 9.1.

Enfin, on a effectué une étude complémentaire où on a comparé la précision atteinte à la précision espérée basée sur la taille de l'échantillon effectivement observé. On a constaté que la fréquence des estimations deux fois moins précises que prévu ou pire passait de 12 % à 5 % pour les cultures, de 9 % à 5 % pour le bétail et de 31 % à 7 % pour les dépenses.

On en conclut qu'en général la précision atteinte est acceptable et diffère de la précision espérée principalement à cause du traitement pour la non-réponse totale. Ceci indique que la plan de sondage est robuste et que la base de sondage pour les dépenses à cause d'un problème dans l'identification des fermes incorporées au Québec et en Ontario lors du recensement et lors de l'enquête. Enfin, on a noté une certaine détérioration de la base de liste, vieille de deux ans déjà lors de l'enquête, et ce, surtout en raison des faillites et des ventes de ferme.

10. CONCLUSION

De façon générale, les résultats de l'enquête ont été sensiblement améliorés suite à l'utilisation du nouveau plan de sondage. Également, la réduction des tailles échantillonnelles a permis de réaliser des économies et de réduire de façon appréciable le fardeau de réponse des fermiers enquêtés. Cependant, certaines difficultés persistent, surtout au niveau des dépenses des fermes incorporées au Québec et en Ontario. Des travaux additionnels sont envisagés pour résoudre ces difficultés.

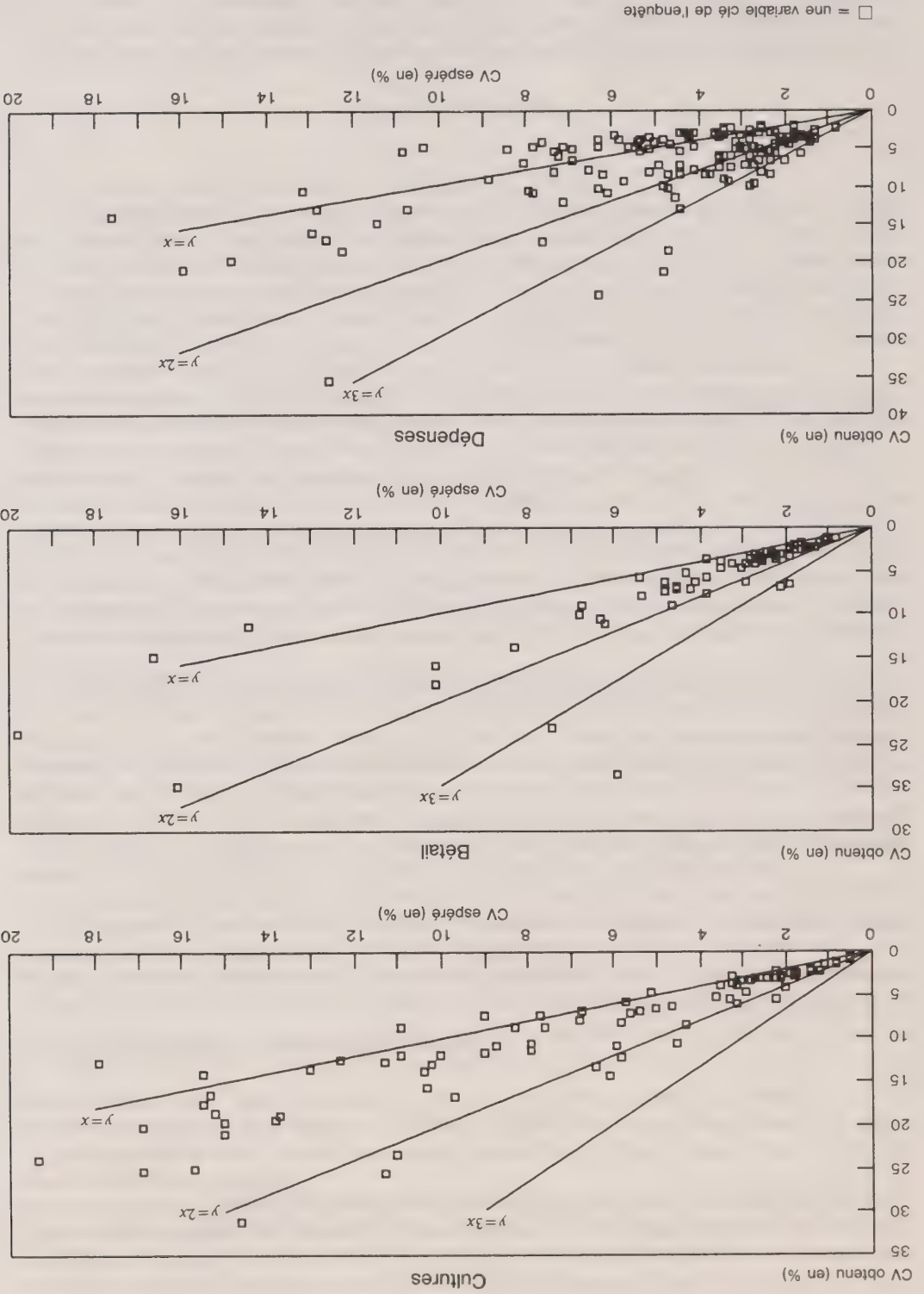


Figure 3. Comparaison de la précision des estimations de variables clés de l'enquête de 1988 à celle qui était espérée lors du développement du plan de sondage.

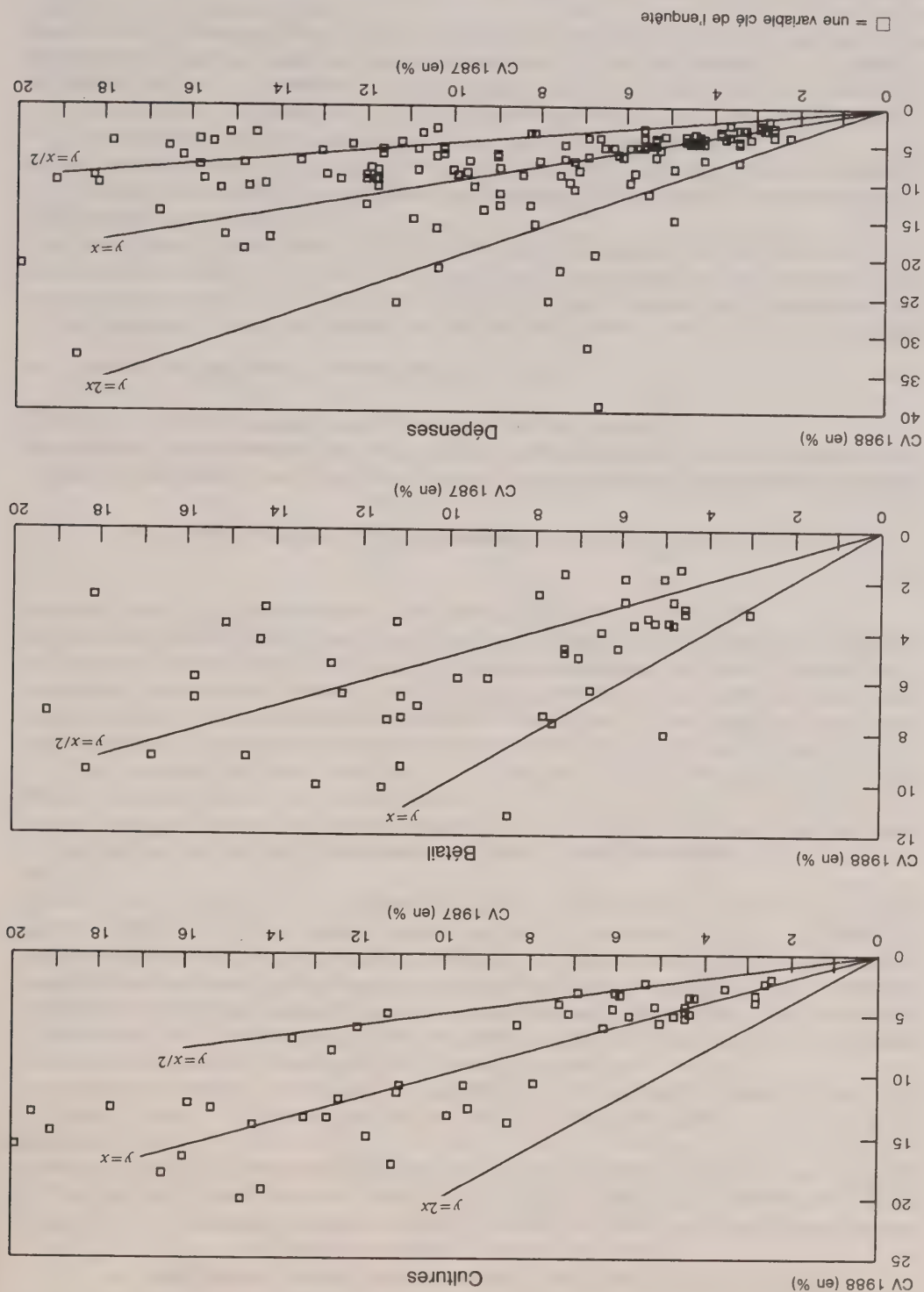


Figure 2. Comparaison de la précision des estimations de variables clés de l'enquête de 1988 à celle de 1987 par catégories de questions.

8. FACTEURS AFFECTANT LA PRÉCISION DES ESTIMATIONS

Pour mieux apprécier les résultats obtenus à la suite de l'enquête de 1988, il est nécessaire d'apporter des précisions sur trois facteurs qui affectent la fiabilité des estimations. Ces facteurs sont la taille de l'échantillon, le traitement de la non-réponse totale et la méthodologie de l'estimation.

D'abord, la taille de l'échantillon pour la liste L1 de la région de la CCB a été réduite de 10% comparativement à l'équivalent de cette liste dans l'ancien plan de sondage. Cette réduction est surtout motivée par le désir de réaliser des économies.

En second lieu, la méthodologie pour traiter la non-réponse totale a été modifiée en 1988. Auparavant, on utilisait les données d'une autre ferme dans la même strate pour imputer des données à une ferme qui n'avait rien répondu à l'enquête. Ces données imputées permettaient alors de compléter l'échantillon à sa taille originale. En 1988, au lieu d'imputer les cas de non-réponse totale, on utilise seulement l'échantillon des répondants et on ajuste à la hausse les facteurs de pondération. Ainsi, l'échantillon effectif est réduit comparativement à l'ancienne méthode.

Lors de l'enquête de 1988, on a observé un taux de non-réponse totale variant entre 2% et 13% selon la province. Au niveau national, ce taux s'établissait à 10%. On présente au tableau 3 les détails des taux de non-réponse.

Le dernier facteur est la méthodologie de l'estimation. Dans les bases de liste, on utilise les estimateurs usuels correspondant à un échantillonnage aléatoire simple stratifié. En ce qui concerne la base aréolaire, on emploie un estimateur décrit dans Wolter (1986 pp. 19-26) et correspondant à un plan de sondage avec répliques indépendantes. Les estimations provinciales sont obtenues en additionnant la contribution des bases de liste et aréolaire car, rappelons-le, ces deux bases sont indépendantes et représentent des domaines mutuellement exclusifs. Les détails de l'estimation se retrouvent dans Lynch (1988).

9. ÉVALUATION DE LA PERFORMANCE DU NOUVEAU PLAN

Pour évaluer la performance du nouveau plan, la précision des estimations obtenues en 1988 est comparée, dans un premier temps, à celle de l'enquête de 1987 et, dans un deuxième temps, à la précision espérée lors du développement du plan de l'enquête.

9.1 Enquête de 1988 contre enquête de 1987

Deux tendances s'opposent lorsque l'on compare la précision des estimations de 1988 à celle des estimations de 1987. D'un côté, les estimations de 1988 devraient être plus précises puisque le plan de sondage de 1987 était vieux de 4 ans déjà. Par contre, les deux facteurs de réduction de la taille de l'échantillon décrits à la section 8 militent en faveur d'une précision moins élevée des estimations de 1988.

La précision est comparée en employant le coefficient de variation (CV) des estimations du niveau provincial provenant de l'union des bases de liste L1 et aréolaire. Ces estimations sont celles de plusieurs variables clés dont le CV en 1987 n'excédait pas 20%.

La comparaison de la précision de 234 estimations est présentée à l'aide de graphiques de la figure 2. Sur ces graphiques, chaque carré représente le CV d'une estimation tel qu'atteint en 1987, en abscisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des variables clés situées à l'intérieur de chaque zone délimitée par les droites $Y = X/2$, $Y = X$ et $Y = 2X$.

On observe que près de 60% des estimations des cultures sont plus précises en 1988 qu'en 1987. Celles qui le sont moins, le sont par peu dans la plupart des cas. Pour le bétail, près de 95% des estimations sont plus précises en 1988. En particulier, 32% des estimations sont même deux fois plus précises. Enfin, plus de 60% des estimations des dépenses sont plus précises

Résultats du plan de sondage des bases de liste

Province	Liste L1			Liste L2		
	N	H	n	n-noyau	N	H
n						

Tableau 1

I.-P.-E.	2830	26	451			
N.-E.	4273	35	550			
N.-B.	3544	39	498			
Québec	41380	80	6096			
Ontario	72598	78	8401			
Manitoba	6712	48	1364	490	18058	29
Saskatchewan	15668	48	3625	1106	45798	41
Alberta	13928	63	2981	909	38504	25
C.-B.(Paix) ^a	494	25	190	190	1187	6
C.-B.(reste) ^b	17042	41	1999			
Total	178469	479	26155	2695	103547	101

^a District de la rivière de la Paix en Colombie-Britannique.
^b Colombie-Britannique, sauf le district de la rivière de la Paix.

Tableau 2

Résultats du plan de sondage aréolaire

Province	N	H	n	n-uniqnes	m
----------	---	---	---	-----------	---

Québec	2065	43	191	182	230
Ontario	2687	49	195	185	259
Manitoba	794	21	277	264	305
Saskatchewan	1496	26	328	308	477
Alberta	1623	32	328	319	434
C.-B.(Paix) ^a	54	7	36	32	58
Total	8719	178	1355	1290	1763

^a District de la rivière de la Paix en Colombie-Britannique.

Tableau 3

Taux de non-réponse totale par province en pourcentage

Province	Refus	Non-contact	Total
I.-P.-E.	0.00	3.55	3.55
N.-E.	0.00	2.18	2.18
N.-B.	0.00	1.61	1.61
Québec	1.71	6.56	8.27
Ontario	2.27	11.11	13.38
Manitoba	3.45	4.03	7.48
Saskatchewan	4.06	6.46	10.52
Alberta	2.68	7.95	10.63
C.-B.	1.78	10.28	12.06
Total	2.32	8.11	10.43

6. MÉTHODES D'ÉCHANTILLONNAGE ARÉOLAIRE

La sélection des échantillons de type aréolaire est basée sur un plan d'échantillonnage stratifié à deux degrés. Les secteurs de dénombrement du recensement et les segments constituent respectivement les unités primaires et secondaires d'échantillonnage.

Étant donné que le plan d'échantillonnage aréolaire n'a pas été modifié au Québec et en Ontario, les paragraphes qui suivent s'appliquent seulement à la région de la CCB.

La première étape consiste à obtenir une mesure de l'activité agricole dans chaque SD de la base en agrégeant au niveau du SD les données des fermes du recensement qui n'appartiennent pas à la liste L1. L'exclusion des fermes de la liste L1 des agrégations produit des distributions de SD qui reflètent fidèlement les caractéristiques des petites fermes. L'emploi subséquent de ces distributions permet de sélectionner un échantillon aréolaire qui complète la liste L1 par rapport aux petites fermes avec une efficacité accrue.

Une fois les agrégations terminées, chaque SD est traité comme une ferme pour les fins d'échantillonnage. La stratégie et les méthodes employées pour la sélection des SD sont très similaires à celles qu'on applique à la liste L1 de la région de la CCB. En effet, on détermine d'abord des SD autoréprésentatifs avec la règle de l'écart sigma. On répartit ensuite, à l'intérieur de régions infraprovinciales, le reste des SD en strates à tirage partiel au moyen de l'algorithme de classification multidimensionnelle CLUSTER. Une classification préliminaire avec FASTCLUS n'est pas nécessaire dans ce cas-ci en raison du nombre relativement peu élevé de SD à traiter, soit jamais plus de 3,000 dans une province. De plus, les transformations appliquées aux variables clés se limitent aux standardisations usuelles. Il n'est pas nécessaire de recourir aux composantes principales car la contribution de la base aréolaire aux estimations provinciales n'est pas suffisante pour justifier une telle approche.

L'allocation aux strates est exécutée avec le même algorithme que celui de la liste et la taille minimale est également fixée à 4. Cela complète, on divise la taille d'échantillon par quatre dans chaque strate et on prélève quatre répliques indépendantes de façon circulaire systématique. Le recours à des répliques facilite le calcul de la variance car il arrive souvent qu'on choisisse une seule unité secondaire par unité primaire.

Une fois les SD sélectionnés, on délimite leur contour sur des cartes topographiques et on parcelise chacun de ceux-ci en segments d'environ 7,5 km² (3 mi²). Ce faisant, on tente, dans la mesure du possible, d'utiliser des limites naturelles comme des routes ou des rivières afin de faciliter ultérieurement le travail des interviewers sur le terrain. Puis, un échantillon aléatoire simple sans remise de segments est prélevé au taux minimum de un sur trente dans chaque SD sélectionné. On note cependant quelques exceptions à la règle: d'abord, on prélève des segments additionnels de façon à ce que le facteur de pondération global n'excède jamais 180; un minimum de deux segments sont choisis dans chacun des SD appartenant aux strates qui font l'objet d'un tirage complet au premier degré; enfin, lorsqu'un même SD figure dans plus d'une réplique, des dispositions sont prises pour éviter le tirage d'un même segment plus d'une fois. Toutes ces exceptions constituent cependant des cas plutôt rares.

7. RÉSULTATS DU PLAN DE SONDAGE

Le tableau 1 contient les résultats du plan de sondage des bases de liste. On y retrouve les quantités suivantes: le nombre de fermes dans la liste (N); le nombre de strates (H); la taille de l'échantillon de fermes (n); et enfin, dans les provinces où cela s'applique, le nombre de fermes dans le sous-échantillon moyen (n-moyau).

Le tableau 2 renferme les résultats du plan de sondage aréolaire dans les provinces où un tel plan est utilisé. On y retrouve les quantités suivantes: le nombre de SD échantillonnés (n); le nombre de strates (H); le nombre de strates où chaque SD est compté une seule fois lorsqu'il apparaît dans plus d'une réplique (n-uni-ques); et enfin le nombre de segments prélevés (m).

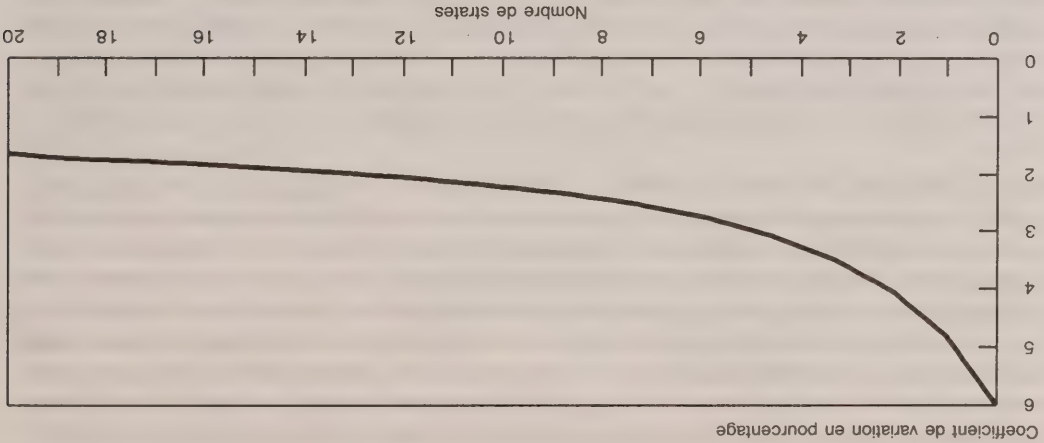


Figure 1. Courbe générale du coefficient de variation en fonction du nombre de strates

Une approche empirique est employée pour décider du nombre de strates. On effectue d'abord plusieurs stratifications et allocations en variant le nombre de strates. Cela fait, on construit la courbe du coefficient de variation obtenu en fonction du nombre de strates, et ce, pour toutes les variables clés et bien d'autres. Ces courbes ont généralement l'allure de la figure 1. On considère que les gains dus à la stratification sont à toutes fins pratiques complétement réalisés au point où la majorité des courbes deviennent quasi horizontales. Le choix du nombre de strates est un compromis entre le point susmentionné et le désir de ne pas former trop de strates pour atténuer les effets des classifications initiales erronées et des changements de strates dans le temps, deux causes importantes d'observations aberrantes.

L'allocation de l'échantillon est multidimensionnelle et utilise généralement les mêmes variables clés que la stratification. L'algorithme d'allocation consiste à minimiser une combinaison linéaire du carré des coefficients de variation des variables clés, sous la contrainte que la taille totale d'échantillon est fixe. En fait, soit c_i le coefficient de variation d'une variable clé, sous la contrainte que $n = n_0$, l'algorithme utilisé est décrit dans Bethel (1986). Cela fait, on procède à des ajustements de façon à ce que la taille minimale soit de 4 et le facteur de pondération maximal soit de 50 dans chaque strate.

Enfin, lorsque l'allocation est complétée, l'échantillon est prélevé dans chaque strate de façon circulaire systématique après avoir ordonné les fermes selon leur région infraprovinciale et leurs dépenses totales d'exploitation. Pour la liste L1 de la région de la CCB, on choisit d'abord l'échantillon complet duquel on sélectionne, toujours de façon circulaire systématique, le sous-

échantillon moyen.

La seconde étape consiste à répartir le reste des fermes de la liste en strates à tirage partiel. Dans la majorité des cas, les strates sont formées à l'intérieur de régions infraprovinciales selon neuf variables clés représentant les trois catégories usuelles: culture, bétail et dépenses. Le nombre de variables dans chaque catégorie est respectivement de un, six et deux.

Le principe sous-jacent à la stratification est le suivant. Chaque ferme est caractérisée par neuf variables et on réunit les fermes qui sont voisines entre elles, le voisinage étant défini en termes de distance euclidienne. Deux algorithmes de classification automatique multidimensionnelle (multivariate clustering) sont employés à cette fin. Ces deux algorithmes seront appelés FASTCLUS et CLUSTER car ils sont disponibles dans les procédures du même nom du progiciel SAS, version 5.

L'algorithme FASTCLUS partitionne un ensemble d'observations en un nombre prédéterminé de grappes mutuellement disjointes. Pour ce faire, l'algorithme choisit d'abord des observations qui servent de noyaux initiaux des grappes. Chaque observation est alors assignée au noyau le plus près et, cela fait, les noyaux sont mis à jour par les moyennes des grappes ainsi formées. Le processus est répété et prend fin lorsque les changements dans les noyaux deviennent petits. Cet algorithme est basé sur les travaux de Hartigan (1975) et MacQueen (1967).

L'algorithme CLUSTER regroupe un ensemble d'observations en grappes mutuellement disjointes de façon hiérarchique. Au départ, chaque observation forme une grappe par elle-même. En utilisant une méthode inspirée de Ward (1963), les deux grappes les plus semblables sont alors agglomérées en une seule et remplacées par cette dernière. Le processus continue et prend fin lorsqu'il ne reste plus qu'une seule grappe. Massart et Kaufman (1983) donnent une introduction à ce genre de classification. Ainsi, on forme autant de partitions de l'ensemble des observations qu'il y a d'observations au départ, et chaque partition correspond à une stratification.

Ces algorithmes sont employés successivement de la façon suivante. On utilise d'abord FASTCLUS pour regrouper les fermes en 250 grappes; ensuite, on fusionne progressivement ces grappes à l'aide de CLUSTER pour finalement former les strates. On effectue une classification préliminaire avec FASTCLUS car l'emploi direct de CLUSTER avec un nombre élevé de dossiers requiert un temps d'ordinateur beaucoup trop important.

Au moment de la stratification, il est impératif que chacune des trois catégories de variables contribue avec le même degré d'importance à la formation des strates. Pour ce faire, on procède à des transformations des variables initiales de stratification. Ces transformations sont exécutées de façon à ce que la somme des variables transformées dans chaque catégorie ait une moyenne de 0 et une variance déterminée, généralement 1. Pour la catégorie culture où il y a une seule variable, il suffit de la standardiser de façon usuelle en lui soustrayant sa moyenne et en divisant par son écart type. Pour les deux autres catégories, on effectue indépendamment dans chacune d'elles deux transformations successives. Soit X_i les variables initiales d'une catégorie donnée C. On effectue d'abord une analyse en composantes principales pour obtenir des variables transformées Y_i . Ces nouvelles variables, de moyenne μ_i et de variance σ_i^2 sont des combinaisons linéaires des anciennes et mutuellement indépendantes. Ensuite, on standardise les Y_i pour obtenir des variables de stratification finales Z_i de la façon suivante:

$$Z_i = \frac{Y_i - \mu_i}{\left(\sum_{i \in C} \sigma_i^2\right)^{1/2}}$$

Ainsi, on constate que $\sum_{i \in C} Z_i$ possède une moyenne de 0 et une variance de 1.

entièrement sauvegardée par la Loi sur la statistique. La seconde différence est qu'il n'est pas nécessaire de sous-échantillonner pour les dépenses car moins de 25 % des fermes de la population enquêtée sont incorporées.

La base areolaire et son plan d'échantillonnage n'ont pratiquement pas été modifiées, faute de ressources, à partir des résultats du dernier recensement. Seules les régions marginales ont été mises à jour, ce qui a résulté en leur agrandissement.

4.3 Provinces Maritimes et le reste de la Colombie-Britannique

Dans chaque province de cette région, le plan de sondage ne comprend qu'une seule base de liste appelée L.1. L'ensemble des fermes du recensement qui font partie de la population enquêtée constitue cette liste L.1. Étant donné qu'une base de liste a tendance à se détériorer avec le temps et qu'il n'y a pas de base areolaire pour la compléter, il devient alors plus difficile de couvrir entièrement la population enquêtée. Cependant, en raison du nombre relativement peu élevé de fermes dans les provinces concernées, soit moins de 30,000, des procédures plus poussées pour tenir à jour la liste ont été mises en place. Ces procédures permettent notamment de détecter des fermes qui ont été oubliées au recensement ou qui ont commencé leurs activités depuis lors. On estime ainsi que la base de liste assure, à toutes fins pratiques, une couverture exhaustive de la population enquêtée.

Dans chaque province, on utilise la même approche qu'au Québec et en Ontario pour stratifier la liste et sélectionner un échantillon de fermes. Cet échantillon est utilisé pour produire toutes les estimations requises.

5. MÉTHODES D'ÉCHANTILLONNAGE DES LISTES

Le prélèvement des échantillons des bases de liste s'appuie sur un plan d'échantillonnage stratifié à un degré où les fermes constituent les unités d'échantillonnage. La stratégie et les méthodes qui sont employées pour le développement de ce plan sont essentiellement les mêmes quelles que soient la province et la liste considérées. Par contre, l'agencement des méthodes et les variables clés en jeu peuvent varier d'un cas à l'autre.

La première étape consiste à identifier les fermes qui ont des caractéristiques distinctes et à procéder à un tirage complet de ces dernières. Ces fermes, dites autoreprésentatives, sont essentiellement de deux types. D'abord, on retrouve celles qui ont une structure d'exploitation unique, soit les pâturages communautaires et les corporations à opérations multiples. En second lieu, il y a les fermes qui se démarquent nettement de la majorité en raison de leur très forte contribution à des variables clés de culture, de bétail et de dépenses. Le dénombrement complet de ces dernières constitue, en raison de l'asymétrie (vers la droite) des distributions traitées, une façon efficace de réduire la variance échantillonnale.

L'identification des fermes à très forte contribution se fait au moyen d'une règle dont les fondements sont intuitifs et qui a donné de bons résultats dans l'ancien plan de sondage de l'enquête. Cette règle, dite de l'écart sigma, est appliquée indépendamment à chaque variable clé sur l'ensemble des fermes ayant une contribution non nulle à la variable en question. Ensuite, sont déclarées autoreprésentatives toutes les fermes dont la contribution est jugée suffisamment élevée selon la règle à l'une ou l'autre des variables clés.

La règle de l'écart sigma adaptée à l'enquête fonctionne de la façon suivante. Soit une distribution unidimensionnelle de points x_i , $i = 1, 2, \dots, N$, $x_i > 0$ pour tout i , et soit σ son écart type; on ordonne la distribution en ordre croissant $x_1 \leq x_2 \leq \dots \leq x_N$; on détermine pour la moitié de la distribution se trouvant à droite de la médiane, la distance entre chaque couple de points successifs $d_i = x_i - x_{i-1}$; soit i_0 le plus petit i pour lequel $d_i \geq \sigma$, alors tous les points $i \geq i_0$ donnent lieu à des fermes autoreprésentatives. Si $d_i < \sigma$ pour tout i , alors aucun point de cette distribution ne se distingue suffisamment des autres pour déclarer une ferme autoreprésentative.

La première base de liste, notée L1, comprend essentiellement les grandes et moyennes fermes du recensement relativement à des variables clés de culture, de bétail et de dépenses. Cette liste est obtenue à l'aide d'un processus itératif qui consiste à établir un seuil pour chaque variable clé et à inclure dans la liste toutes les fermes qui excèdent au moins un de ces seuils. On ajuste indépendamment à la hausse ou à la baisse chacun des seuils de façon à ce que la liste L1 représente, une fois complétée, environ 35 % des fermes de la population enquêtée et de 50 % à 90 % de l'activité agricole totale selon la variable clé considérée. Ces pourcentages sont retenus car l'expérience a démontré que la liste qui en résulte est composée de fermes qui sont individuellement plus stables dans le temps que le reste des fermes de la population enquêtée. Cette stabilité permet de créer des strates qui demeurent homogènes au fil des ans, ce qui est un facteur de conservation d'efficacité du plan d'échantillonnage.

Dans chaque province, la liste L1 est ensuite stratifiée à l'intérieur de régions provinciales selon neuf variables clés. Un échantillon de fermes est sélectionné pour obtenir des données sur les cultures et le bétail. Les données sur les dépenses étant plus difficiles et dispendieuses à recueillir, seul un sous-échantillon, appelé noyau, est tenu de fournir ces renseignements.

La deuxième base de liste, notée L2, renferme les fermes du recensement de plus de 20 acres qui n'ont pas été retenues dans la liste L1. Sa stratification se fait à l'intérieur des districts agricoles selon une seule variable clé, soit la superficie cultivée au recensement. La liste L2 sert à compléter la liste L1 pour les données préliminaires sur les cultures. Ces données doivent être recueillies dans des délais très courts et la base aréolaire ne peut, pour des raisons opérationnelles, satisfaire ces délais.

La base aréolaire comprend tous les secteurs de dénombrement agricoles, sauf ceux qui correspondent aux réserves indiennes et aux régions dites marginales, c'est-à-dire là où il y a peu d'activité agricole. Ces régions marginales comprennent surtout le nord des provinces et les zones en bordure des villes. Les rares fermes du recensement qui sont situées dans les régions marginales sont ajoutées à la liste L1 car seule cette liste est utilisée pour recueillir des renseignements sur toutes les variables de l'enquête.

La base aréolaire est stratifiée en utilisant les mêmes régions provinciales et variables clés que la liste L1. Elle engendre ultimement un échantillon de segments qui sont délimités sur des cartes topographiques. L'identité des fermiers qui exploitent des terrains dans ces segments est obtenue par un dénombrement sur place. Ensuite, des appartements manuels sur noms et adresses permettent de détecter les cas de chevauchement entre les fermes de segments et l'une ou l'autre des bases de liste. Cette détection est essentielle car chaque fois que la base aréolaire est appelée à compléter une base de liste, seules les fermes de segments qui ne chevauchent pas la liste en question sont utilisées. Ainsi, on s'assure que les bases de listes et aréolaire représentent des domaines mutuellement exclusifs.

Des renseignements complets sont exigés de toutes les fermes de segments, sauf celles qui chevauchent la liste L1 car les données pour la liste L1 proviennent de l'échantillon tiré de cette liste.

4.2 Québec et Ontario

Dans chacune de ces deux provinces, on a recours à une seule base de liste, appelée L1, et à une base aréolaire.

La base de liste est composée de l'ensemble des fermes du recensement qui appartiennent à la population enquêtée. La méthodologie employée pour échantillonner cette liste est similaire à celle de la liste L1 de la région de la CCB, à deux différences près. La première différence consiste à séparer des autres les fermes incorporées, c'est-à-dire constituées en sociétés par actions, puis à former indépendamment des strates dans chacun de ces deux groupes. Cette séparation préliminaire est effectuée car seules les fermes incorporées doivent rapporter leurs dépenses dans l'enquête, les dépenses des autres fermes étant obtenues à partir des dossiers fiscaux de Revenu Canada. Il convient de noter que la confidentialité de ces dossiers est

3. POPULATIONS CIBLE ET ENQUÊTE

La population cible comprend toutes les fermes des provinces enquêtées dont la vente de produits agricoles s'est chiffrée à 250 dollars ou plus au cours des 12 mois précédant le début de l'enquête. Faut également partie de la population cible les fermes qui ne satisfont pas au critère du 250 dollars en date de l'enquête mais qui anticipent réaliser au moins cette somme au cours des 12 mois suivant l'enquête. Ces dernières sont relativement peu nombreuses; ce sont des fermes qui ont débuté leurs activités juste avant l'enquête ou qui sont temporairement inactives.

La population enquêtée, c'est-à-dire celle qui est effectivement échantillonnée, exclut les fermes exploitées par les institutions ainsi que les fermes situées dans les réserves ou établissements indiens. Les termes institution, réserve indienne et établissement indien sont définis dans Statistique Canada (1987, pp. 115-117, 145, 152). Le rapport coûts-bénéfices associé à la collecte des données pour ces types de fermes est très élevé. Ainsi, on les exclut afin de permettre une meilleure utilisation des ressources consacrées à l'enquête. La contribution des exclusions à la production agricole nationale est faible et on l'estime en utilisant des facteurs d'ajustements calculés à partir des données du recensement.

4. BASES DE SONDAGE ET LEUR UTILISATION

En théorie, la population enquêtée se répartit en deux groupes, le premier regroupant les fermes qui ont été dénombrées au recensement et le second toutes les autres fermes. Ces autres fermes correspondent au sous-dénombrement du recensement et aux fermes dites nouvelles, c'est-à-dire celles dont l'exploitation a débuté après le recensement.

Le premier groupe est couvert, en tout ou en partie selon la province, par une ou deux bases de liste formées à même le registre des fermes du recensement. Pour compléter les bases de liste et assurer une couverture complète de la population enquêtée, on a recours à une base aréolaire qui est créée à partir des secteurs de dénombrement (SD) agricoles. Par secteur de dénombrement, on entend la région géographique qui est dénombrée par un agent recenseur; de plus, un secteur est dit agricole s'il renferme au moins une ferme au recensement. Le recours à une base aréolaire est nécessaire afin de pallier aux lacunes des bases de liste, notamment en ce qui concerne leurs difficultés à détecter les nouvelles fermes.

Les exigences de l'enquête en matière d'estimation et les caractéristiques de l'agriculture canadienne varient selon la région. Pour mieux tenir compte de ces variations, on divise le territoire couvert par l'enquête en trois régions et on utilise un plan de sondage distinct dans chacune d'entre elles. Les trois régions concernées sont les suivantes: les provinces des Prairies et le district de la Paix en Colombie-Britannique; le Québec et l'Ontario; et enfin les provinces Maritimes et le reste de la Colombie-Britannique. La première de ces régions est appelée région de la Commission canadienne du blé (CCB) car c'est le territoire auquel s'étend la juridiction de cet organisme.

La taille totale des échantillons dans chacune des trois régions est établie essentiellement à partir du budget global disponible pour la collecte des données. À l'intérieur de chaque région, la répartition des échantillons entre les diverses provinces et, selon le cas, entre les diverses bases, dépend de plusieurs facteurs. Les principaux facteurs en jeu sont la règle de la racine carrée de la taille de la population enquêtée, les répartitions historiques de l'enquête et les résultats de diverses analyses portant sur la précision espérée des estimations.

4.1 Région de la Commission canadienne du blé

Dans cette partie du Canada, on utilise deux bases de liste et une base aréolaire dans chaque province.

Le plan de sondage de l'enquête nationale sur les fermes de 1988

C. JULIEN et F. MARANDA¹

RÉSUMÉ

L'Enquête nationale sur les fermes est une enquête par échantillonnage qui produit des estimations annuelles sur une variété de sujets reliés à l'agriculture canadienne. En 1988, l'enquête a été dotée d'un nouveau plan de sondage. Ce nouveau plan fait intervenir des bases de sondage multiples et des méthodes d'échantillonnage multidimensionnelles qui sont différentes de celles du plan précédent. Dans cet article, on décrit d'abord la stratégie et les méthodes utilisées pour le nouveau plan de sondage. Ensuite, on apporte des précisions sur quelques facteurs qui affectent la précision des estimations. Enfin, on évalue la performance du nouveau plan suite à son utilisation.

MOTS CLÉS: Échantillonnage à buts multiples; base multiple; base aréolaire; stratification multivariée.

1. INTRODUCTION

L'Enquête nationale sur les fermes est une enquête par échantillonnage probabiliste portant sur plusieurs sujets reliés à l'agriculture canadienne. Elle est menée annuellement en juin et juillet dans toutes les provinces à l'exception de Terre-Neuve où une enquête séparée est effectuée. L'ancien plan de sondage de l'enquête datait de 1983 et il était basé sur les résultats du Recensement de l'agriculture de 1981. On en retrouve une description dans Ingram et Davidson (1983). Or, depuis 1981, la population des exploitations agricoles a subi plusieurs changements qui ont entraîné une perte d'efficacité de ce plan. De plus, les exigences de l'enquête ont quelque peu changé au cours des ans et il était devenu impératif d'apporter des corrections aux échantillons.

Pour ces raisons, on a développé un nouveau plan de sondage. Ce nouveau plan s'appuie sur les résultats du Recensement de l'agriculture de 1986 et il est devenu opérationnel à l'été 1988.

2. OBJECTIFS DE L'ENQUÊTE

L'objectif premier de l'enquête consiste à produire des estimations actuelles et fiables reflétant les niveaux et les tendances annuelles de plus d'une centaine de variables agricoles. Ces variables se répartissent essentiellement en trois catégories: les superficies ensemencées de l'année en cours; la taille des cheptels au premier juillet; et enfin les recettes et dépenses d'exploitation pour l'année civile précédente. En termes de fiabilité, l'enquête vise des coefficients de variation inférieurs à 5 % pour les paramètres importants à l'échelle provinciale.

Les données de l'enquête sont normalement agrégées au niveau des provinces. Cependant, surtout pour des fins analytiques, l'enquête produit également des résultats pour des régions infraprovinciales au moyen de méthodes d'estimation par le domaine.

Un autre objectif important de l'enquête est de procurer un échantillon maître duquel on choisit des sous-échantillons servant à d'autres enquêtes agricoles effectuées par Statistique Canada.

¹ C. Julien est méthodologiste, Section de la qualité des données et de l'analyse du recensement, Division des méthodes des d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6; F. Maranda est chef, Section des méthodes d'enquêtes agricoles, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

- VARDEMAN, S., et MEEDEN, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.
- WHITE, D. (1987). Mean squared error of estimators using two stage sampling for stratification and prior information. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_{k=1}^K N_k S_k^2 + \frac{1}{1-f} \sum_{j=1}^J N_j S_j^2. \quad (18)$$

La validité de cette approximation repose sur trois constatations. Premièrement, lorsqu'on applique l'équation (18) dans les cinq exemples pour lesquels il existe des données simulées, on obtient des résultats très comparables. L'erreur type approximative calculée à l'aide de l'équation (12) est 113.25, 108.97, 108.09, 106.77 et 106.32 pour $f' = .10, .20, .30, .40$ et $.50$ respectivement. Ces valeurs se rapprochent sensiblement de celles indiquées dans le tableau 3 sous la rubrique $M = 0$ de la colonne $ET(\hat{\tau})$ avec $m = 500$ ou $m = 2500$. Deuxièmement, après avoir analysé l'erreur engendrée par chaque approximation, on a constaté que cette erreur était négligeable (sauf peut-être en ce qui a trait à l'approximation A6) pour des populations et des échantillons relativement grands. En ce qui concerne l'approximation A6, la loi des grands nombres indique néanmoins que n_j pourra être remplacée adéquatement par son espérance mathématique si l'échantillon est suffisamment grand. Enfin, comme nous le décrivons ci-dessous, l'équation (18) se ramène à l'équation exacte pour les trois plans d'échantillonnage courants. Premièrement, nous sommes en présence d'un échantillonnage stratifié ordinaire (selon les strates de premier niveau) lorsque $J = K$, $P_j = P_k$ pour $j = k$, et $c = 1$. L'équation (18) se réduit alors à $\text{var}(\hat{\tau}) \approx (1-f)/f \sum_{k=1}^K N_k S_k^2$, qui est bien l'équation exacte. Par ailleurs, nous sommes en présence d'un sondage à deux phases ordinaire pour stratification lorsque $K = 1$ et là encore, l'équation (18) se ramène à l'équation exacte (voir Cochran, 1977, p.329). De même, nous sommes en présence d'un échantillonnage stratifié ordinaire selon les strates de second niveau lorsque $f = 1$ (K et la stratification a priori ne sont plus vraiment pertinents dans ce cas) et une fois de plus, l'équation (18) se réduit à l'équation exacte.

BIBLIOGRAPHIE

- ALLEN, J.R. (1984). Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS) in the United States. *Seminars in Oncology*, 11, 4-11.
- CASAREALE, D. *et al.* (1984/5). Prevalence of AIDS-associated retrovirus and antibodies among male homosexuals at risk for AIDS in Greenwich Village. *AIDS Research*, 1, 407-421.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.
- HAN, C. (1973). Double sampling with partial information on auxiliary variables. *Journal of the American Statistical Association*, 68, 914-918.
- HANSEN, M.H., et HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.
- KURITSKY, J.N. *et al.* (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion*, 26, 205-207.
- LUI, K. *et al.* (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immuno-deficiency syndrome. *Proceedings of the National Academy of Sciences*, 83, 3051-3055.
- RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- REDFIELD, R.R., et BURKE, D.S. (1987). Shadow on the land: the epidemiology of HIV infection. *Viral Immunology*, 1, 69-81.
- ROBERT-GUROFF, M. (1986). Prevalence of antibodies to HTLV-I, -II, and -III in intravenous drug abusers from an AIDS endemic region. *Journal of the American Medical Association*, 255, 3133-3137.

Nous obtenons ensuite

(14)
$$\text{var}(t | s') = \frac{1 - c}{n_j} \sum_{j=1}^f \frac{f_2 c}{n_j} \left[\sum_{k=1}^k (n_{kj}' - 1) s_{kj}^2 + \sum_{k=1}^k n_{kj}' (y_{kj}' - \bar{y}_{j'})^2 \right].$$

Les seconde et troisième approximations consistent à substituer un (1) à $n_j' / (n_j' - 1)$ (A2) et n_{kj}' à $(n_{kj}' - 1)$ (A3) dans l'équation (14). En calculant l'espérance du premier terme de l'équation (14), nous obtenons

(15)
$$E \left[\frac{1 - c}{n_j} \sum_{j=1}^f \sum_{k=1}^k n_{kj}' s_{kj}^2 \right] \approx \frac{1 - c}{n_j} \sum_{j=1}^f \sum_{k=1}^k n_{kj} S_{kj}^2.$$

Une approximation additionnelle (A4) est nécessaire dans l'équation ci-dessus; nous ne tenons pas compte de la possibilité que $n_{kj}' \leq 1$ pour tout k, j . Nous devons aussi calculer l'espérance du second terme de l'équation (14). La formule exacte pour calculer cette espérance est

(16)
$$\frac{1 - c}{n_j} \sum_{j=1}^f \sum_{k=1}^k n_{kj} (y_{kj} - \bar{y}_{j'})^2 + a_1 \sum_{k=1}^k S_{kj}^2 - a_2 \left\{ \right.$$

où $a_1 = 1 - f - E[n_{kj}'(1 - n_{kj}'/N_{kj})/n_j]$ et $a_2 = E[(\sum_{k=1}^k n_{kj}'(y_{kj} - \bar{y}_{j'})^2)/n_j]$. Nous remarquons en premier lieu que $|a_1| \leq 1$ et qu'on peut ne pas en tenir compte lorsqu'il est combiné à N_{kj} dans l'équation (15) (approximation A5). En outre, si dans a_2 , on remplace n_j' par son espérance mathématique, $f N_j$ (approximation A6), étant donné que $E[\sum_{k=1}^k n_{kj}'(y_{kj} - \bar{y}_{j'})^2] = 0$, nous avons

$$a_2 \approx \frac{1}{f N_j} \text{var} \left(\sum_{k=1}^k n_{kj}' (y_{kj} - \bar{y}_{j'}) \right) \approx (1 - f) \sum_{k=1}^k \frac{N_j}{N_{kj}} (1 - W_{kj}) (y_{kj} - \bar{y}_{j'})^2$$

où nous avons finalement substitué N_k à $(N_k - 1)$ (approximation A7) en calculant la variance de la variable hypergéométrique n_{kj}' . Lorsque nous comparons a_2 au terme analogue de l'équation (16) affecté du coefficient N_{kj} , nous constatons que a_2 est en soi à peu près négligeable. Enfin, si nous ne tenons pas compte des différences entre N_{kj} et $(N_{kj} - 1)$ ou entre N_j et $(N_j - 1)$ (approximation A8), nous pouvons combiner les équations (15) et (16) pour obtenir

$$E(\text{var}(t | s')) \approx \frac{1 - c}{N_j} \sum_{j=1}^f \frac{f c}{N_j} \frac{1}{N_j} \sum_{k=1}^k [(N_{kj} - 1) S_{kj}^2 + N_{kj} (y_{kj} - \bar{y}_{j'})^2]$$

$$= \frac{1}{N_j} \sum_{j=1}^f \frac{f c}{S_{kj}^2}.$$

(17)

En combinant les équations (13) et (17), nous obtenons finalement

5. APPLICATIONS

La méthode de stratification duale que nous venons de décrire peut être utilisée à deux niveaux. À un premier niveau, on peut recourir à l'échantillonnage double avec strates préliminaires sans utiliser d'information préalable sur les tailles ou les moyennes de strates. À un second niveau, plus complexe, pourvu que l'on dispose d'information préalable sur le nombre d'unités dans chaque strate de second niveau qui proviennent de chacune des strates de premier niveau et que l'on connaisse le degré de fiabilité de cette information, il sera possible de réduire davantage l'erreur type de l'estimateur grâce à cette information.

Cette méthode de sondage à deux phases et la méthode d'estimation correspondante pourraient être utilisées dans l'enquête nationale que l'on propose pour évaluer la propagation du VIH (syndrome d'immunodéficience acquise). La période d'incubation de la maladie, que l'on estime à 4,5 ans en moyenne (Lui et coll., 1986), rend cette enquête d'autant plus nécessaire; en revanche, elle peut être très difficile à réaliser à cause de facteurs psychosociaux et financiers. C'est pourquoi il est nécessaire de rechercher des méthodes qui permettent de réduire la taille de l'échantillon tout en conservant le même niveau de précision.

Allen (1984) présente des données qui permettent de croire qu'il est possible de répartir la population des États-Unis selon une série de facteurs qui peuvent servir à définir des catégories de risque. Parmi les facteurs connus, qui peuvent servir à définir les strates de premier niveau, notons l'âge, le sexe, la présence de certaines maladies, la nationalité, le statut d'immigrant et le lieu géographique. Parmi les facteurs inconnus, qui peuvent être déterminés au moyen d'une interview, notons l'orientation sexuelle et la toxicomanie. Les données sur la prévalence du VIH dans divers sous-groupes peuvent aussi bien être incluses dans l'estimation globale de la prévalence que servir à déterminer la répartition de l'échantillon. On dispose de données de ce genre pour les donneurs de sang par exemple (Kuritsky et coll., 1986), les militaires (Redfield et Burke, 1987), les toxicomanes consommateurs de drogues injectables de Queens, New York (Robert-Guioff et coll., 1986) et les homosexuels de Greenwich Village (Casarale et coll., 1984/1985). Bien que cette information préalable puisse servir à réduire le coût d'échantillonnage et à accroître le niveau de précision, le problème de la confidentialité des tests ne se pose pas avec moins d'acuité; du reste, il faudra étudier ce problème sous tous ses angles avant de pouvoir tirer des résultats probants de cette étude.

REMERCIEMENTS

L'auteur tient à exprimer sa gratitude à l'arbitre pour les précieux commentaires que celui-ci lui a transmis en ce qui concerne l'échantillonnage non proportionnel.

ANNEXE

Calcul de l'espérance et de la variance sans information préalable (échantillonnage proportionnel)

À l'aide de la notation définie dans la section 2, nous calculons tout d'abord $E(\tau)$. L'espérance conditionnelle, étant donné s' , est $E(\tau | s') = 1/f \sum_j n_{ij} y_{ij}$. En exprimant $n_{ij} y_{ij}$ par $\sum_k n_{kj} y_{kj}$, nous obtenons $E(\tau) = E(E(\tau | s')) = 1/f \sum_j \sum_k E(n_{kj} y_{kj} | n_{kj}) = 1/f \sum_j \sum_k E(n_{kj}) Y_{kj} = \tau$ puisque n_{kj} est hypergéométrique, étant donné la traction de sondage f et N_{kj} unités contenues dans la strate de premier niveau k et la strate de second niveau j . Par conséquent, τ est non biaisé en l'occurrence (abstraction faite de l'approximation A1). Le calcul de la variance s'effectue selon le même raisonnement sauf qu'il est beaucoup plus détaillé. Nous ne retiendrons ici que certaines étapes du calcul, plus particulièrement celles où il est question d'approximation. Tout d'abord, en appliquant les deux arguments de condition définis plus haut, nous obtenons

(8)
$$+ \frac{N_{k.}(n_{k.}-1)}{n_{k.}} \sum^j \frac{n_{k.}}{n_{kj.}} \left(y_{kj.} - \sum^j \frac{n_{k.}}{n_{kj.}} y_{kj.} \right)^2$$

Lorsque nous avons examiné le cas de l'échantillonnage proportionnel au début de cette section, nous avons proposé deux estimateurs pour τ , un fondé sur un échantillon de seconde phase combiné et l'autre, sur un échantillon non combiné. Dans les deux cas, il s'agissait d'un estimateur sans biais; de plus, lorsque $f_{k.} = f$ pour tous k et $c_{kj} = c$ pour tous k et tous j , l'équation (7) se ramène à l'équation (3), c'est-à-dire à la formule de la variance approximative de l'estimateur fondé sur un échantillon de seconde phase combiné.

Enfin, nous servant une fois de plus des résultats de Rao, nous déterminons un plan optimal pour répartir les ressources consacrées à l'échantillonnage. Supposons que nous disposions de D dollars pour les deux phases du sondage; il en coûte d_{kj} dollars pour échantillonner une unité dans P_{kj} à la phase 1 et d_j dollars pour échantillonner une unité dans P_j à la phase 2. Étant donné ces coûts, nous voulons déterminer les valeurs $f_{k.}$ et c_{kj} qui minimisent la variance de $\hat{\tau}$. En utilisant l'inégalité de Cauchy pour l'échantillon de la phase 2 dans chaque strate de premier niveau, nous constatons que, quelle que soit la valeur de $f_{k.}$, la fraction de sondage pour la strate de second niveau j est définie par l'expression

(9)
$$c_{kj} = S_{kj} \left(\frac{d_j (S_{k.}^2 - \sum^j w_{kj} S_{kj}^2)}{d_{kj}} \right)^{1/2}.$$

Par ailleurs, le coût prévu effectif (pour les deux phases d'échantillonnage) pour chaque unité tirée de la strate de premier niveau k à la phase 1 est

(10)
$$d_{k.}^{(e)} = d_{k.} + \sum^j w_{kj} c_{kj} d_j.$$

De ce point de vue, on peut considérer la première phase du sondage comme un échantillonnage stratifié ordinaire, où le coût (effectif) du prélèvement de l'unité dans $P_{k.}$ est déterminé à l'aide de l'équation (10). Cochran (1977, p.97) définit la formule de répartition voulue pour la première phase:

(11)
$$\frac{n_{k.}'}{n_{k.}} = \frac{\sum^{k'} \frac{N_{k'}. S_{k'}. / \sqrt{d_{k'}^{(e)}}}{N_{k.} S_{k.} / \sqrt{d_{k.}^{(e)}}}$$

où

(12)
$$n' = \sum^k n_{k.}' = D \sum^k \frac{N_{k.} S_{k.} / \sqrt{d_{k.}^{(e)}}}{\sum^{k'} \frac{N_{k'}. S_{k'}. / \sqrt{d_{k'}^{(e)}}}.$$

Grâce aux modifications proposées par Rao, il existe désormais une solution pour les cas où l'un ou plusieurs des c_{kj} seraient supérieurs à un. On peut aussi modifier les résultats de la façon habituelle afin de réduire au maximum le coût d'échantillonnage lorsque la variance est connue.

$$D_a = \frac{D - d_0}{N} = f(d_1 + cd_2). \quad (4)$$

Les paramètres f et c devant satisfaire l'équation (4), nous cherchons à minimiser l'équation (3), $\text{var}(\hat{\tau})$, maintenant définie

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{1-c} T_1 + \frac{fc}{1-c} T_2. \quad (5)$$

Nous en arrivons assez facilement à la solution suivante:

$$c = \left[\frac{d_1 T_2}{d_2 (T_1 - T_2)} \right]^{1/2} \quad (6)$$

f étant calculé à l'aide de l'équation (4). Si $T_1 \leq T_2$, nous choisissons automatiquement $c = 1$ puisque dans les circonstances, la stratification a priori est plus efficace que la stratification a posteriori.

Dans le cas de l'échantillonnage non proportionnel, l'estimateur donné est biaisé et le calcul du biais et de l'erreur type de cet estimateur est très long. Toutefois, en modifiant légèrement le plan d'échantillonnage de la seconde phase et, du même coup, l'estimateur $\hat{\tau}$, on obtient un estimateur sans biais. Après avoir décrit les modifications requises, nous calculons la variance de l'estimateur ainsi qu'un estimateur non biaisé de cette variance et nous déterminons une méthode optimale pour répartir les ressources affectées à l'échantillonnage entre les strates de premier niveau et de second niveau.

La modification du plan d'échantillonnage consiste à effectuer l'échantillonnage de la seconde phase à l'intérieur des strates de premier niveau, plutôt que grouper les éléments des strates de premier niveau par strate de second niveau. Ainsi, étant donné n_{kj} unités qui appartiennent à $s' \cap P_{kj}$, nous avons une fonction $v_{kj}(\cdot)$ (comme $v_j(\cdot)$) définie dans la section 2) qui appelle un échantillon aléatoire simple de taille $n_{kj} = v_{kj}(n_{kj}) = c_{kj} n_{kj}$ qui doit être prélevé dans $s' \cap P_{kj}$. À partir de cet échantillon, nous calculons les valeurs \hat{y}_{kj} et \hat{s}_{kj}^2 , qui ont été définies dans la section 2. La formule de l'estimateur est maintenant $\hat{\tau} = \sum_k 1/f_k \cdot \sum_j n_{kj} \hat{y}_{kj}$. Or, comme les échantillons (et par voie de conséquence les estimateurs) sont indépendants d'une strate préliminaire à l'autre, $\hat{\tau}$ est la somme des estimateurs indépendants des K totaux de strates de premier niveau, où chaque estimateur repose sur un sondage à deux phases ordinaire. Ainsi, les résultats de Rao (1973) s'appliquent à chaque strate de premier niveau et nous constatons en premier lieu que $\hat{\tau}$ est non biaisé parce que ses termes sont des estimateurs non biaisés des totaux de strates de premier niveau correspondants. En deuxième lieu, si nous nous servons des résultats de Rao, nous avons

$$\text{var}(\hat{\tau}) = \sum_k \frac{1}{f_k} \left[(N_{k\cdot} - n_{k\cdot}) S_{k\cdot}^2 + \sum_j N_{kj} S_{kj}^2 (1/c_{kj} - 1) \right]. \quad (7)$$

En outre, un estimateur non biaisé de $\text{var}(\hat{\tau})$ est défini par l'expression:

$$\widehat{\text{var}}(\hat{\tau}) = \sum_k N_{k\cdot} \left[(N_{k\cdot} - 1) \sum_j \left(\frac{n_{k\cdot} - 1}{n_{k\cdot}} - \frac{N_{k\cdot} - 1}{n_{kj} - 1} \right) \frac{n_{kj} S_{kj}^2}{n_{k\cdot} n_{kj}} \right]$$

En résumé, s'il est possible d'appliquer une constante de pondération à des éléments d'information préalable sur la répartition des unités entre les strates, on peut réduire sensiblement l'erreur type de l'estimateur grâce aux méthodes décrites ci-dessus. Même si on ne peut trouver de constante de pondération ou qu'on ne dispose pas de l'information préalable voulue, il est encore possible de réduire l'erreur type de l'estimateur par la stratification duale en choisissant $M = 0$ si l'information préalable sur W_{kj} est maigre ou inexistante, ou $M = \infty$ si l'information préalable est exacte. Cela démontre l'importance du cas où $M = 0$. Celui-ci est analysé en détail dans la section suivante.

4. BIAIS, ERREUR TYPE ET RÉPARTITION OPTIMALE SANS INFORMATION PRÉALABLE

Lorsqu'il n'y a pas d'information préalable, nous posons $M_j = 0$ et $\tilde{M}_k = 0$ pour chaque $1 \leq j \leq J$, et chaque $1 \leq k \leq K$. Dans cette section, nous supposons aussi au départ que l'échantillon, dans les deux phases, est proportionnel à la taille du groupe d'où il est tiré, c'est-à-dire pour chaque k , $n_k = fN_k$. $(c, a, d, f)_k = f$, pour tous k) et pour chaque j , $n_j = cn_j(c, a, d, v_j(x) = cx$, pour tous j). Cela implique automatiquement une approximation (designée ci-dessous par A1) puisque la taille des échantillons ainsi obtenus n'est pas nécessairement un nombre entier. Toutefois, pour des populations assez grandes et des fractions de sondage f et c suffisamment élevées, cette approximation a peu d'effet sur les calculs qui vont suivre.

Dans le cas qui nous occupe, $\hat{\mu}_{.j}$ se ramène à $\bar{y}_{.j}$ et Π_{kj} à n'_{kj}/n'_k ; par conséquent, nous avons $\hat{\tau} = 1/f \sum_{j=1}^J n'_j \bar{y}_{.j}$. Le calcul de l'espérance mathématique et de la variance de $\hat{\tau}$ est décrit en annexe. Dans cette annexe, nous posons deux arguments de condition; nous posons tout d'abord une condition en fonction de s' puisque l'échantillon de la seconde phase est une fonction de s' puis, étant donné la nature hypergéométrique multidimensionnelle de l'échantillon de la première phase, une autre condition en fonction de n'_{kj} , qui est le nombre d'unités de s' tirées de la strate préliminaire k , qui appartiennent aussi à la strate de second niveau j .

Nous montrons tout d'abord que l'estimateur $\hat{\tau}$ est non biaisé (abstraction faite de l'approximation A1) et que sa variance peut être calculée approximativement par la formule suivante:

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{1-c} \sum_k N_k S_k^2 + \frac{fc}{1-c} \sum_j N_j S_j^2. \quad (3)$$

Comme nous le voyons plus en détail dans l'annexe, l'équation (3) produit des valeurs propres des valeurs simulées, repose sur des approximations dont l'erreur est faible pour de grandes populations et des échantillons suffisamment grands, et correspond à l'équation exacte dans les trois cas courants. En outre, on peut montrer facilement que la variance calculée à l'aide de cette équation est toujours inférieure à celle calculée pour le sondage à deux phases ordinaire.

Comme dans n'importe quel modèle de stratification, il faut penser à un plan optimal. Il s'agit ici de déterminer une variance minimum pour un coût donné. À cette fin, posons $T_1 = \sum_k N_k S_k^2$ et $T_2 = \sum_j N_j S_j^2$. Pour les besoins de la cause, nous supposons que ces valeurs sont connues. En réalité, nous ne disposons bien sûr que d'estimations. Dans un deuxième temps, définissons D comme le budget total, d_0 comme les frais de mise en marche, d_1 comme le coût unitaire pour l'échantillon de la première phase et d_2 comme le coût unitaire pour l'échantillon de la seconde phase. Si nous définissons D_a comme le montant (en dollars) consacré à l'échantillonnage par unité de population, nous avons

Tableau 2 Principaux résultats des simulations répétées $m = 500, f = .10$ et $c = 1.0$

Simulation #	M_0	$ET(\hat{\tau})$		
		$M = 0$	$M = m$	$M = M_0$
1	600	113.55	109.67	109.62
2	700	113.42	109.50	109.45
3	700	113.92	109.86	109.78
4	600	113.61	109.71	109.66
5	600	113.56	109.74	109.70
				112.00
				111.80
				112.07
				112.17

Principales caractéristiques et de $ET(\hat{\tau})$ comme fonction de M									
m	f'	c	$ET(\hat{\tau}_2)$	M_0	$ET(\hat{\tau})$			Réduction relative de $ET(\hat{\tau})$ en pourcentage	$M = 0 \quad M = m \quad M = \infty$
					$M = M_0$	$M = 0$	$M = m$		
500	.10	1.00	126.29	600	109.62	113.55	109.67	112.00	-3.6
500	.20	.50	115.19	600	107.95	109.02	107.97	110.72	54.5
500	.30	.33	111.80	600	107.87	108.25	107.87	110.38	56.1
500	.40	.25	109.22	750	106.51	106.76	106.52	108.29	65.6
500	.50	.20	107.98	700	106.17	106.28	106.18	107.55	67.7
2500	.10	1.00	126.29	*	≤ 106.20	113.33	106.42	106.20	-0.8
2500	.20	.50	115.19	*	≤ 105.76	108.67	106.02	105.76	59.0
2500	.30	.33	111.80	*	≤ 106.63	108.18	106.87	106.63	57.2
2500	.40	.25	109.22	*	≤ 105.77	106.59	105.94	105.77	70.1
2500	.50	.20	107.98	*	≤ 105.81	106.34	105.96	105.81	80.5
* -- > 10,000									

qui figurent dans la colonne "Réduction relative de $ET(\hat{\tau})$ " sont calculées au moyen de la formule $100 [\min(ET(\hat{\tau}_2), 113.27) - ET(\hat{\tau})] / [\min(ET(\hat{\tau}_2), 113.27) - 105.47]$.

Le tableau 3 nous permet de faire plusieurs constatations importantes. Premièrement, lorsque $m = 500, M_0$ se rapproche sensiblement de cette valeur tout en demeurant légèrement supérieur. C'est ce que prévoit White (1987) lorsqu'il analyse le cas où $K = 1$. Pour $m = 2500$, malgré que $M_0 > 10,000$ dans chaque cas, on remarque que $ET(\hat{\tau})$ à $M = m$ se rapproche sensiblement de l'erreur type minimum à $M = M_0$.

Deuxièmement, lorsque $M = m$, la réduction relative de $ET(\hat{\tau})$ varie de 46% à plus de 90%. En outre, lorsque $M = 0$, ce qui correspond à la stratification duale sans information préalable sur aucune caractéristique de la population, la réduction relative de $ET(\hat{\tau})$ est toujours supérieure à 50% sauf pour ce qui a trait à la plus petite fraction de sondage de première phase, $f = .10$. Lorsque, pour $f = .10$, il n'y a pas d'information préalable et que l'échantillon de la première phase est petit, il est préférable d'utiliser les strates préliminaires et de ne pas tenir compte des strates réelles. En revanche, si l'on dispose d'un ensemble d'estimations préliminaires pour W_{kj} mais qu'on ne sait trop quels poids attribuer à ces valeurs, on peut recourir à la stratification a posteriori habituelle, où $M = \infty$. Si l'information préalable est juste (en l'occurrence, $m = 2500$), la réduction relative de $ET(\hat{\tau})$ est toujours supérieure à 80%. Même lorsque l'information préalable n'est qu'à demi exacte (en l'occurrence, $m = 500$), la réduction relative de l'erreur type se situe entre 16 et 33%.

préliminaire de taille m ($= 500$ ou 2500) prélevé dans chaque strate de premier niveau. En d'autres termes Π_{kj} est défini comme la proportion des m unités de la strate de premier niveau k qui appartiennent aussi à la strate de second niveau j .

En deuxième lieu, nous avons défini les constantes de pondération comme suit pour chaque simulation: $M_1 = M_2 = M$ pour tous $M \in \{0, 100, 200, 300, \dots, 10\,000, \infty\}$ Il convient de rappeler que $M = \infty$ correspond au modèle de stratification a posteriori habituel, où on n'utilise pas du tout les données de l'échantillon courant pour estimer la taille des groupes. En troisième lieu, l'échantillon de la première phase est stratifié suivant les strates préliminaires, les fractions de sondage f'_j étant définies $f'_1 = f'_2 = f, f \in \{.10, .20, .30, .40, .50\}$. Rappelons que dans cette phase d'échantillonnage, seules les données relatives aux strates de second niveau sont observées. La collecte de ces données est une opération vraisemblablement peu coûteuse.

En revanche, l'échantillonnage d'une unité à la phase 2, où l'on détermine la présence du virus, est une opération qui est supposée assez coûteuse. Les unités échantillonnées forment un sous-échantillon de l'échantillon de la première phase, stratifié suivant les strates de second niveau. Là encore, les fractions de sondage sont définies comme étant égales ($v_j(n'_j) = [c_j n'_j]$ pour une valeur n'_j suffisamment grande, et $c_1 = c_2 = c_3 = c$) et pour permettre la comparaison de simulations, on choisit la valeur de c de manière que la proportion de la population totale incluse dans l'échantillon de la phase 2 soit toujours la même ($.10$).

Le processus suivant est répété $R = 50\,000$ fois: formation d'un échantillon préliminaire de taille m à partir duquel sont établies des estimations préliminaires Π_{kj} de W_{kj} . Ensuite, formation d'un échantillon stratifié suivant les strates de premier niveau avec fractions de sondage f . Seules les données relatives aux strates de second niveau sont observées. Ensuite, formation d'un sous-échantillon stratifié suivant les strates de second niveau avec fractions de sondage c et classement des unités de ce sous-échantillon en deux groupes: infectées et non infectées. Finalement, calcul de $\hat{\tau}$ pour chaque valeur de M considérée. L'erreur type de $\hat{\tau}$ est estimée au moyen des R valeurs simulées de $\hat{\tau}$. Rappelons toutefois que dans la réalité, l'erreur type d'une valeur estimée dépendra des valeurs de Π_{kj} utilisées. Dans la présente simulation, ces valeurs diffèrent à chaque itération et il vaut mieux ici considérer l'erreur type estimée comme une moyenne à long terme pour diverses distributions de $\hat{\tau}$ combinées suivant la distribution des valeurs Π_{kj} fondées sur l'échantillon préliminaire.

Les simulations ont été exécutées sur un ordinateur IBM3031. Pour cet exemple, où $y_i \in \{0, 1\}$ pour tout i , toutes les quantités aléatoires sont des fonctions de variables hypergéométriques ou de variables hypergéométriques multidimensionnelles indépendantes. Compte tenu de ce que la distribution conditionnelle d'un élément d'une distribution hypergéométrique multidimensionnelle, étant donné n l'importe quel sous-ensemble des autres coordonnées, est elle-même hypergéométrique, toutes les quantités aléatoires ont été simulées au moyen du sous-programme de simulation hypergéométrique GCHPR du IMSL 92DP. Pour la première combinaison de valeurs de m et f (500 et $.10$), on a répété la simulation cinq fois afin de vérifier la cohérence. On retrouve dans les tableaux 2 et 3 des caractéristiques pertinentes de la variation de l'ET($\hat{\tau}$) simulée; dans le premier cas, il s'agit de la variation en fonction de M pour les cinq simulations répétées et dans le second cas, de la variation en fonction de M pour des simulations faites pour diverses valeurs de f et de m . Le tableau 2 ne donne que les résultats saillants de la simulation, qui confirment la cohérence et indiquent que le nombre d'itérations choisis est suffisamment élevé. Le tableau 3 permet d'établir une comparaison entre d'une part, les meilleures méthodes en usage actuellement (sondage à deux phases ordinaire ou échantillonnage stratifié suivant les strates de premier niveau) et d'autre part, la méthode idéale, c'est-à-dire celle où les strates réelles sont connues. L'erreur type d'un estimateur fondé sur un échantillon stratifié suivant les strates de premier niveau (strates préliminaires) seulement est de 113.27 tandis que celle d'un estimateur fondé sur un échantillon stratifié suivant les strates réelles est de 105.47 . Par conséquent, si nous désignons par $\hat{\tau}_2$ l'estimateur fondé sur le sondage à deux phases ordinaire et que nous reconnaissons que ET($\hat{\tau}_2$) varie en fonction de f et de c , les valeurs

En nous servant des estimations préliminaires, des estimations de l'échantillon courant et des poids de fiabilité, nous estimons W_{kj} et $\hat{Y}_{.j}$ par $\hat{\Pi}_{kj} = (\tilde{M}_k \cdot \Pi_{kj} + n_{kj}) / (\tilde{M}_k + n_{kj}^k)$ et $\hat{\mu}_{.j} = (M_{.j} \mu_{.j} + n_{.j} \hat{y}_{.j}) / (M_{.j} + n_{.j})$ respectivement. Enfin, nous construisons un estimateur $\hat{\tau}$ du total de population τ en remplaçant n l'importe quelle quantité non observée de l'équation (1) par la valeur estimée correspondante, calculée ci-dessus. Ainsi, nous avons

$$\hat{\tau} = \sum_{j=1}^J \left\{ n_{.j} \hat{y}_{.j} + (n_{.j} - n_{.j}) \hat{\mu}_{.j} + \sum_{k=1}^K (N_k - n_k^k) \hat{\Pi}_{kj} \hat{\mu}_{.j} \right\} \quad (2)$$

Vardeman et Meeden ne traitent pas le calcul du biais et de la variance de $\hat{\tau}$ pour le cas général. White (1987) examine le cas où $K = 1$ et $M_{.j} = 0, 1 \leq j \leq J$. Avant d'analyser des cas plus complexes, nous allons examiner les résultats d'une simulation appliquée à une population hypothétique.

3. ETUDE DE MONTE CARLO

Pour les besoins de cette simulation, nous nous servons d'une population et d'un plan d'échantillonnage particuliers pour estimer, comme dans l'exemple du début, le taux de propagation d'une maladie infectieuse. Pour une population de 10,000 personnes qui sont prédisposées à cette maladie, on suppose que la prévalence est plus élevée parmi les 5,000 personnes qui demeurent dans la partie ouest de la région étudiée. Par conséquent, nous divisons la population en $K = 2$ strates (régions est et ouest). Ensuite, nous supposons que le sondeur est en mesure, grâce à de l'information supplémentaire obtenue facilement, de classer les personnes en fonction du degré de risque de contracter la maladie. Voir le tableau 1 pour les données détaillées de la population.

Pour l'estimation du nombre total de personnes infectées ($\tau = 2302$), nous supposons qu'il n'existe aucune information préalable sur les proportions de strates $Y_{.1}$, $Y_{.2}$ et $Y_{.3}$ et nous posons par conséquent $M_{.1} = M_{.2} = M_{.3} = 0$. Nous devons considérer quatre autres éléments essentiels pour l'estimation: 1) les estimations préliminaires $\{\Pi_{kj}; k = 1, 2; j = 1, 2, 3\}$ concernant la répartition des personnes entre les strates de premier niveau et de second niveau, 2) les constantes de pondération \tilde{M}_1 et \tilde{M}_2 attribuées à ces estimations préliminaires, 3) le plan d'échantillonnage et l'échantillon effectif de la première phase et 4) le plan d'échantillonnage et l'échantillon effectif de la seconde phase. Ces quatre éléments sont exposés en détail ci-dessous.

En premier lieu, White (1987) a observé qu'une solution efficace pour $K = 1$ était de choisir une constante de pondération M égale à la taille de l'échantillon sur lequel était fondée l'information antérieure. Compte tenu de cette observation, nous avons prévu, pour chaque simulation, la formation d'un ensemble d'estimations préliminaires $\{\Pi_{kj}\}$ à partir d'un échantillon

Lieu de résidence	Groupe de risque j	1 Faible	2 Moyen	3 Élevé	Total
Région est ($k = 1$)	40/4000	80/800	100/200	220/5000	
Région ouest ($k = 2$)	2/200	80/800	2000/4000	2082/5000	
Total	42/4200	160/1600	2100/4200	2302/10000	

Tableau 1
Nombre de cas de maladies/taille du groupe pour les strates de premier et de second niveau

$$S^2_k = \frac{1}{i \in P_k} \sum (y_i - \bar{Y}_k)^2.$$

Nous pouvons aussi écrire l'équation suivante:

$$(1) \qquad \qquad \qquad \tau = \sum^j N_{.j} X_{.j}.$$

Enfin, soit $W_{kj} = N_{kj}/N_k$, c.-à-d. que W_{kj} est la proportion d'unités de la strate préliminaire k qui appartiennent à la strate j .

Voyons maintenant la méthode d'échantillonnage. Dans la première phase, on tire sans remise un échantillon aléatoire simple stratifié s' ; cet échantillon est constitué de n'_k unités (fraction de sondage de la première phase: $f'_k = n'_k/N_k$) prélevées dans la strate préliminaire k . Les échantillons prélevés dans des strates préliminaires différentes sont indépendants. Pour ces $n' = \sum_k n'_k$ unités, on observe des strates de second niveau j_i . Suivant la notation définie plus haut, n'_{kj} désigne le nombre d'unités de s' tirées de la strate préliminaire k , qui appartiennent aussi à la strate de second niveau j . Par ailleurs, $n'_j = \sum_k n'_{kj}$ est le nombre total d'unités de s' qui font partie de la strate de second niveau j . Cet ensemble est désigné par s'_j . Ces quantités sont connues tandis que celles qui reposent sur les valeurs y , comme y' et s'^2 (avec les quatre genres d'indices), ne le sont toujours pas. Pour les besoins de cet article, la moyenne d'un ensemble vide est définie comme nulle et la variance s'^2 d'un groupe dont l'effectif est un ou zéro est aussi définie comme nulle. Notons que pour $1 \leq k \leq K$, les vecteurs aléatoires (n'_k, \dots, n'_{kj}) sont indépendants et suivent chacun une distribution hypergéométrique multidimensionnelle.

Pour la seconde phase d'échantillonnage, nous décomposons s' en $\cup_{j=1}^J s'_j$, c.-à-d. que nous procédons à une stratification a posteriori. Pour chaque j , posons $v_j(\cdot)$ comme une application connue dans et sur les nombres entiers positifs, où $v_j(0) = 0$ et $1 \leq v_j(x) \leq x$ si $x \geq 1$. L'échantillon de la seconde phase s est aussi stratifié mais il est un sous-ensemble de s' et est stratifié en fonction de la stratification a posteriori. L'échantillon prélevé dans s'_j est désigné par s_j et sa taille est $n_j \equiv v_j(n'_j)$. À ce stade-ci, nous pouvons observer des valeurs y ainsi que les paramètres établis à l'aide de ces valeurs, y_j et s_j^2 , qui sont la moyenne et la variance de population finie des unités de la strate j dans l'échantillon de la seconde phase.

Lorsqu'on calcule les valeurs estimées de τ selon la formule de Vardeman et Meeden, on peut faire intervenir des estimations préliminaires des tailles relatives de strates pour chaque strate préliminaire et des moyennes de strates. Ici, nous disposons d'estimations préliminaires pour les valeurs W_{kj} et X_j et ces estimations sont désignées par Π_{kj} et μ_j , respectivement. Dans l'estimateur défini plus bas, ces estimations sont affectées de constantes de pondération qui reflètent le rapport entre le degré de fiabilité de l'estimation et celui de l'information correspondante tirée de l'échantillon courant. On désigne par $\tilde{M}_k \in [0, \infty]$ le degré de certitude attribué à l'ensemble $(\Pi_{k1}, \dots, \Pi_{kJ})$ pour chaque k et par $M_j \in [0, \infty]$ le degré de certitude attribué à μ_j , pour chaque j . Dans l'échantillon courant, l'ensemble (W_{k1}, \dots, W_{kJ}) est estimé par $(n'_{k1}/n'_k, \dots, n'_{kJ}/n'_k)$ et repose sur un échantillon aléatoire simple de taille n'_k . Par conséquent, le rapport entre la valeur de \tilde{M}_k et celle de n'_k par exemple reflètera le rapport entre le degré de fiabilité de Π_{kj} et celui de n'_{kj}/n'_k . De même, dans l'échantillon courant, X_j est estimé par y_j et repose sur un échantillon de taille n_j ; le rapport entre les valeurs de M_j et de n_j reflète donc, là aussi, le rapport entre le degré de fiabilité de l'estimation préliminaire et celui de l'estimation tirée de l'échantillon courant. Un poids de fiabilité (ou degré de fiabilité) nul signifie que l'on n'utilise pas l'information préalable et un poids égal à l'infini signifie, comme pour l'utilisation des tailles de strates dans le modèle de stratification a posteriori habituel, que l'on ne se sert pas de l'information correspondante de l'échantillon courant.

membres de chaque strate. Dans ce dernier cas, la population est stratifiée selon divers facteurs, dont certains sont connus et d'autres non; ces derniers peuvent toutefois être déterminés sans trop de frais dans une première phase de sondage.

Par exemple, nous pouvons évaluer la propagation d'une maladie infectieuse en ayant recours à l'échantillonnage. Si le dépistage de la maladie est une opération coûteuse, il est sage de recourir à la stratification (par catégorie de risque) pour réduire la taille de l'échantillon de la seconde phase. Le sexe, l'âge, le lieu de résidence, l'origine ethnique, les habitudes d'hygiène et le contact avec des porteurs de la maladie sont tous des facteurs qui peuvent déterminer les catégories de risque. Comme certains de ces facteurs ne sont pas connus avant l'échantillonnage, on peut utiliser le modèle de Vardeman et Meeden puisqu'on peut déduire les véritables catégories de risque à l'aide des facteurs connus.

Prenons comme autre exemple le sondage à deux phases pour non-réponse. Si nous appliquons une version élargie de la méthode de Hansen et Hurwitz (1946), nous avons une population qui est subdivisée en deux strates: les répondants et les non-répondants. Les méthodes que nous analysons ici s'appliquent lorsqu'on dispose d'information dans des strates de premier niveau, lesquelles servent ensuite à déterminer si une unité est susceptible ou non de faire partie du groupe des répondants.

L'utilisation d'information préalable dans les plans de sondage à deux phases est un sujet qui a déjà été traité dans des ouvrages de l'échantillonnage. Han (1973), par exemple, s'est servi de l'information préalable sur une variable de régression auxiliaire (qui devait être mesurée dans un échantillon de première phase) pour construire une hypothèse simple (disons H_0) concernant la moyenne de cette variable. Il a ensuite testé H_0 à l'aide des observations de l'échantillon de première phase. Si l'hypothèse H_0 était vérifiée, il se servait de la valeur définie par celle-ci H_0 dans l'estimateur; dans le cas contraire, il se servait de la moyenne de l'échantillon. L'utilisation du premier estimateur de Vardeman et Meeden (information globale seulement) fait l'objet d'une analyse dans White (1987). Celui-ci s'attache à optimiser le choix des constantes de pondération pour l'information préalable par rapport à l'information contenue dans l'échantillon courant. En ce qui concerne le présent article, nous envisageons le cas où il existe aussi de l'information préalable sur les membres des strates. Après avoir présenté la notation pertinente dans la section 2, nous analysons un exemple simulé dans la section 3. Par la suite, nous analysons des estimateurs non biaisés au point de vue de la variance, de l'estimation non biaisée de cette variance et de la répartition optimale des ressources affectées à l'échantillonnage dans deux contextes différents. Enfin dans la section 5, nous voyons des applications de ces méthodes.

2. MODÈLE DE POPULATION ET PLAN DE SONDAGE

Nous allons maintenant exposer le modèle de population et le plan de sondage proposé. Soit une population finie P d'unités identifiées $1, 2, \dots, N$ et les valeurs inconnues correspondantes y_1, y_2, \dots, y_N . Désignons le total de population par $\tau = \sum_{i=1}^N y_i$. Pour $1 \leq i \leq N$, l'unité i appartient à la strate de second niveau j , $1 \leq j \leq J$, que nous ne connaissons pas, et appartient en même temps à la strate de premier niveau (strate préliminaire) k , $1 \leq k \leq K$, que nous connaissons.

Il y a des paramètres de population qui exigent une notation spéciale. Ces paramètres sont pertinents: l'absence d'indice sert à désigner la population totale, " $K \dots$ " renvoie à la strate préliminaire k , $1 \leq k \leq K$, " $J \dots$ " renvoie à la strate de second niveau j , $1 \leq j \leq J$, et l'indice " kj " sert à désigner l'intersection de la strate préliminaire k et de la strate de second niveau j . Les symboles N_j , \bar{Y}_j , et S_j^2 représentent le nombre d'éléments, la moyenne et la variance de population finie respectivement. Enfin, désignons par P , P_k , P_j et P_{kj} les sous-ensembles de P qui correspondent aux quatre groupes définis ci-dessus. Par exemple, nous avons

Estimation au moyen d'un échantillonnage double et d'une stratification duale

DONALD B. WHITE¹

RÉSUMÉ

Nous cherchons ici à estimer le total d'une population finie qui est stratifiée à deux niveaux: un niveau secondaire, caractérisé par une faible variabilité intrastate mais inconnu avant la première phase du sondage, et un niveau préliminaire (stratification a priori), qui permet de prévoir avec assez d'efficacité, pour chaque unité, les résultats de la stratification a posteriori. Dans un sondage à deux phases visant à tenir compte de la non-réponse, par exemple, la stratification a posteriori peut servir à diviser une sous-population en deux groupes: répondants et non-répondants. Nous appliquons les estimateurs de Vardeman et Meeden (1984) suivant diverses hypothèses concernant le genre d'information préalable utilisée. Au moyen d'une simulation, nous analysons l'erreur type de ces estimateurs en regard de celle des méthodes courantes. Lorsqu'il n'existe pas d'information préalable et que l'on a recours à l'échantillonnage proportionnel, l'estimateur est non biaisé et sa variance est calculée par approximation. Dans ce cas, la variance est toujours inférieure à celle de l'échantillonnage double ordinaire pour stratification. Par ailleurs, lorsqu'il n'y a pas d'information préalable mais que l'on a recours à l'échantillonnage non proportionnel, il est possible, moyennant une légère modification du plan d'échantillonnage de la seconde phase, de déterminer un estimateur non biaisé de même que la variance correspondante, un estimateur non biaisé de cette variance et un mode de répartition optimale pour les deux phases du sondage. Enfin, nous examinons des applications de ces méthodes.

MOTS CLÉS: Sondage à deux phases; information préalable; estimation de la variance; répartition optimale; non-réponse.

1. INTRODUCTION

Les plans d'échantillonnage stratifié n'utilisent pas tous le même genre d'information préalable. Par exemple, le modèle de stratification habituel suppose que nous disposons de toute l'information voulue sur les membres de chaque strate. La stratification a posteriori est utile lorsque nous disposons d'information globale sur les tailles de strates mais que nous n'avons aucune information sur les unités proprement dites. Par ailleurs, l'échantillonnage double pour stratification suppose qu'il n'existe aucune information préalable sur les strates. En outre, il est nécessaire d'avoir une certaine connaissance des valeurs de la population si l'on veut, par exemple, répartir entre les strates les ressources consacrées à l'échantillonnage (voir, par exemple, Cochran, 1977, p. 96-99 et 331-332). Les hypothèses rigoureuses qui sont propres à ces plans de sondage et à ces modèles de population ne sont pas toujours vérifiées à cause de la divergence qui peut exister entre la population à l'étude et l'information préalable (parfois périmée). Cherchant à résoudre cet écart, Vardeman et Meeden (1984) ont défini deux estimateurs qui combinent les données relatives aux membres des strates et aux tailles et moyennes de strates avec les données correspondantes pour l'échantillon courant. Les deux estimateurs sont utilisés dans des circonstances fondamentalement différentes. Le premier est utilisé lorsque l'information préalable est uniquement globale, c'est-à-dire lorsque elle ne porte que sur les tailles et les moyennes de strates, tandis que le second estimateur est utilisé lorsqu'on dispose en plus de données partielles sur les

¹ Donald B. White, Département de statistique State University of New York, Buffalo 249 Farber Hall Buffalo, New York 14214.

ou

$$A = \sum_H [n_h / (n_h - 1)] \left[\left\{ \sum_{j=1}^f e_{.hj}^2 \right\} - e_{.h.}^2 / n_h \right],$$

$$B = \sum_D \left\{ (1 - m_d / M_d) [1 / (m_d - 1)] \right\} \cdot$$

$$\left(\sum_H [n_h / (n_h - 1)] \left\{ \sum_{j=1}^f e_{dhj}^2 \right\} - e_{d.}^2 / n_h \right) \cdot$$

$$C = - \sum_H f_h n_h / (n_h - 1) \left[\sum_{j=1}^f \{ e_{.hj}^2 - \text{var}_{2hj} \} - \{ e_{.h.}^2 - \text{var}_{2h} \} / n_h \right],$$

$f_h = n_h / N_h$ est la fraction de sondage de la première phase pour la strate h , et var_{2hj} et var_{2h} sont définis par l'équation (6).

Notons que si toutes les fractions de sondage de la première phase sont très petites, C a un effet négligeable dans l'équation (8). De toute manière, le pire qui puisse arriver si on laisse tomber C est que var sera entaché d'un biais vers le haut, puisque $E(C) \leq 0$.

Notons en outre que var se réduirait à A si, en plus d'avoir un C suffisamment petit pour qu'on puisse ne pas en tenir compte; on avait comme plan de sondage un plan classique à deux degrés d'échantillonnage; autrement dit, si chaque domaine se trouvait entièrement dans une des UPE échantillonnées au départ, de sorte que $y_{d.} = y_{dh.} = y_{dh.}$ et $B = 0$. Ce fait ne devrait pas surprendre, car A est l'estimateur type de la variance dans un plan d'échantillonnage à deux degrés lorsque la méthode d'échantillonnage utilisée au premier degré est l'éas avec remise (Cochran 1977, p. 307). Lorsque les fractions de sondage du premier degré sont négligeables, la distinction entre l'éas avec et sans remise se dissipe.

La partie droite de l'équation (8) peut, en principe, être négative. C'est que B est souvent négatif (puisque $y_{d.} \geq y_{dh.} \geq y_{dh.}$), tandis que A peut théoriquement être aussi petit que zéro. Kott et Johnston (1988) ont appliqué une formule semblable à (6) à des données recueillies au moyen d'une enquête du département américain de l'Agriculture. Dans les 41 cas qu'ils ont examinés, la valeur absolue de B était toujours inférieure à 7 % de A .

Dernière observation: Lorsque $B \leq 0$ et comme $E(C) \leq 0$, on obtient, en prenant A seul, une estimation modérée, non ambiguë et non négative la $\text{var}(X)$.

BIBLIOGRAPHIE

- COCHRAN, R., et HUDDLESTON, H. (1969). Unbiased estimates for stratified subsample designs. U.S. Department of Agriculture, Statistical Reporting Service.
- COCHRAN, R., et HUDDLESTON, H. (1970). Unbiased estimates for stratified subsample design. *Proceedings of the Section on Social Statistics, American Statistical Association*, 265-267.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3ième ed.). New York: Wiley.
- KOTT, P.S., et JOHNSTON, R. (1988). Estimating the non-overlap variance component for multiple frame agricultural surveys. RAD Staff Report No. SRB-NERS-8805, U.S. Department of Agriculture, National Agricultural Statistics Service.
- SÄRNDAAL, C.E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

où E_k et var_k expriment, l'espérance et la variance relatives à la k^{e} phase d'échantillonnage.

On appelle souvent le premier terme de l'équation (2) la variance de la première phase parce qu'il représente la variance qu'on obtiendrait si toutes les USB appartenant à une UPE étaient échantillonnées faisaient partie du sous-échantillon.

On appelle souvent le deuxième terme de l'équation (2) la variance de la seconde phase. Celle-ci est plus facile à estimer que la variance de la première phase, aussi allons-nous nous y attaquer en premier. L'estimation de la variance de la première phase pose un problème, en ce sens qu'on ne peut estimer la valeur totale qui nous intéresse pour une UPE dans l'échantillon de première phase qu'en ayant recours au sous-échantillon. Or, il est bien connu que si l'on met le total estimé pour une UPE au lieu du total réel dans la formule habituelle pour l'estimation de la variance d'un échantillon à une phase, on obtient un estimateur biaisé.

3.1 Estimation de la variance de la seconde phase

Pour tout échantillon original, un estimateur non biaisé de $\text{var}_2(X)$ est automatiquement un estimateur non biaisé de $E_1[\text{var}_2(X)]$. Pour le démontrer, supposons que v_2 est un estimateur non biaisé de $\text{var}_2(X)$ quel que soit l'échantillon. Etant donné que $E_2[v_2 - \text{var}_2(X)] = 0$ pour chaque S^1 possible, l'espérance de $E_2[v_2 - \text{var}_2(X)]$ pour la première phase doit aussi être égale à zéro. Par conséquent, $E(v_2) = E_1 E_2(v_2) = E_1[\text{var}_2(X)]$.

Etant donné notre S^1 en particulier,

$$\text{var}_2 = \sum_D (1 - m_d/M_d) [m_d/(m_d - 1)] \left\{ \sum_{i \in R^d} e_i^2 \right\} - e_{d..}^2/m_d \quad (3)$$

est l'estimateur non biaisé classique pour $\text{var}_2(X)$. En outre, l'équation (3) serait valide quel que soit l'échantillon de première phase obtenu. Il en découle que var_2 est aussi un estimateur non biaisé pour $E_1[\text{var}_2(X)]$.

3.2 Estimation de la variance de la première phase

Prenons une UPE j dans la strate h . La valeur $e_{.hj}$ est un estimateur non biaisé de (N_h/n_h) fois la valeur totale dans toutes les USB de l'UPE j , que celles-ci soient dans le sous-échantillon actuel ou non. Par conséquent, $E_2(e_{.hj})$ est exactement égal à (N_h/n_h) fois la valeur totale dans toute les USB de l'UPE j . Compte tenu de cela, nous obtenons donc par la formule suivante un estimateur non biaisé de la variance de X pour la première phase:

$$\text{var}_1[E_2(X)] =$$

$$\sum_{h=1}^H (1 - n_h/N_h) [n_h/(n_h - 1)] \left[\sum_{n_h}^j \{E_2(e_{.hj})\}^2 - \{E_2(e_{.h.})\}^2/n_h \right]. \quad (4)$$

Telle quelle, l'équation (4) est de peu d'utilité, car elle suppose que nous connaissons la valeur des $\{E_2(e_{.hj})\}^2$ et des $\{E_2(e_{.h.})\}^2$. Elle laisse néanmoins penser qu'on pourrait estimer $\text{var}_1[E_2(X)]$ sans biais si l'on pouvait trouver des estimateurs non biaisés pour les termes $\{E_2(e_{.hj})\}^2$ et $\{E_2(e_{.h.})\}^2$ que l'on pourrait insérer dans l'équation (4).

Remarquez tout d'abord que $e_{.hj}^2$ et $e_{.h.}^2$ ne sont pas des estimateurs non biaisés de $\{E_2(e_{.hj})\}^2$ et de $\{E_2(e_{.h.})\}^2$. En fait,

$$E_2(e_{.hj}^2) = \{E_2(e_{.hj})\}^2 + \text{var}_2(e_{.hj}),$$

$$E_2(e_{.h.}^2) = \{E_2(e_{.h.})\}^2 + \text{var}_2(e_{.h.}).$$

tandis que

(5)

Estimons uniquement le total pour un élément qui nous intéresse en particulier. À cette fin,

soit,

S^1 = l'ensemble d'USE appartenant à une UPE tirée à la première phase d'échantillonnage, que ces USE soient dans le sous-échantillon ou non,

S_{hj} = l'ensemble d'USE prélevées dans l'UPE j de la strate h ,

S_h = l'ensemble d'USE prélevées dans la strate h ,

R_d = l'ensemble d'USE prélevées dans le domaine d ,

x_i = la valeur qui nous intéresse pour l'USE i ,

$e_i = (N_h/n_h)(M_d/m_d)x_i$ (en supposant que $i \in S_h \cap R_d$) la valeur qui nous intéresse, étendue à l'ensemble, pour l'USE i ,

$$e_{dhj} = \sum_{i \in S_{hj} \cap R_d} e_i,$$

$$e_{dh} = \sum_{i \in S_h \cap R_d} e_i,$$

$$e_{d\cdot\cdot} = \sum_{i \in R_d} e_i,$$

$$e_{\cdot hj} = \sum_{i \in S_{hj}} e_i, \text{ et}$$

$$e_{\cdot h} = \sum_{i \in S_h} e_i.$$

Notons que lorsque S_{hj} est vide, e_{dhj} et $e_{\cdot hj}$ égalent zéro. De même, lorsque S_h est vide, e_{dh} et $e_{\cdot h}$ égalent zéro, et lorsque R_d est vide, e_{dhj} , e_{dh} , et $e_{d\cdot\cdot}$ égalent zéro.

La formule suivante est un estimateur non biaisé de X , la somme des valeurs de x_i pour l'ensemble des USE de la population:

$$X = \sum_D \sum_{i \in R_d} e_i. \tag{1}$$

Pour le constater, notons que $\bar{X} = \sum_{i \in S^1} (N^D/n^D)x_i$ est un estimateur non biaisé de X pour la première phase d'échantillonnage, tandis que \bar{X} est un estimateur non biaisé de \bar{X} pour la deuxième phase d'échantillonnage. Mathématiquement, $E_1(\bar{X}) = X$ et $E_2(\bar{X}) = \bar{X}$, ce qui implique que $E(\bar{X}) = E_1E_2(\bar{X}) = X$.

3. VARIANCE DE \bar{X}

La consultation de n'importe quel ouvrage sur la théorie de l'échantillonnage (p. ex. Cochran 1977, p. 276), nous apprend que la variance d'un estimateur pour échantillon à deux phases comme \bar{X} est

$$\text{var}(\bar{X}) = \text{var}_1[E_2(\bar{X})] + E_1[\text{var}_2(\bar{X})], \tag{2}$$

Estimation de la variance lorsque l'échantillon aréolaire de première phase est restreint

PHILLIP S. KOTT¹

RÉSUMÉ

Dans cet article, l'auteur propose une formule d'estimation non biaisée de la variance pour un plan de sondage à deux phases tel qu'utilisé dans de nombreuses enquêtes agricoles. Ce genre de plan consiste, dans un premier temps, à prélever des unités primaires d'échantillonnage (UPB) définies selon des critères géographiques au moyen d'une méthode d'échantillonnage aléatoire simple stratifié; les unités secondaires d'échantillonnage dans les UPB prélevées sont ensuite stratifiées à leur tour en fonction de leurs caractéristiques, et l'on prélève un échantillon de ces dernières au cours d'une deuxième phase d'échantillonnage aléatoire simple stratifié.

MOTS CLÉS: Échantillon à deux phases; unité primaire d'échantillonnage; unité secondaire d'échantillonnage; non biaisé.

1. INTRODUCTION

Supposons que nous prélevons dans une base aréolaire stratifiée un échantillon d'unités primaires d'échantillonnage (UPB) définies selon des critères géographiques. Chaque UPB échantillonnée contient un certain nombre d'unités secondaires d'échantillonnage (USE) qui sont stratifiées à leur tour en fonction de leurs caractéristiques. On tire alors un échantillon d'USE dans chaque nouvelle strate. Pour éviter toute confusion, nous désignerons les premières strates aréolaires par le nom de strates et les nouvelles strates formées en fonction des caractéristiques des USE, par le nom de *domaines*. À chacune des deux phases du plan de sondage, nous procédons par échantillonnage aléatoire simple (éas) stratifié sans remise.

Dans cet article, nous proposons une formule d'estimation non biaisée de la variance pour le plan de sondage décrit ci-dessus, lequel est utilisé dans un grand nombre d'enquêtes agricoles (voir par exemple Kott et Johnston 1988), mais n'est pas limité à ce genre d'enquête. Cette formule est une généralisation d'un plan proposé par Cochran et Huddleston (1969, 1970), qui comportait un échantillonnage aléatoire simple (éas) non stratifié à la première phase d'échantillonnage. Elle est aussi un cas particulier d'une formule d'estimation de la variance de Särndal et Swensson (1987). La formule de Särndal et Swensson (leur équation 4.4) exige le calcul d'une probabilité d'inclusion conjointe pour chaque paire d'USE sous-échantillonnées. Cette façon de procéder n'est pas commode pour l'application à l'étude, dans laquelle il faut considérer six situations distinctes (selon que les deux USE proviennent ou non de la même UPB, de la même strate et/ou du même domaine). La démarche présentée ici est le fait d'un raisonnement complètement différent.

2. DÉFINITIONS

Supposons que nous partons d'une base de sondage aréolaire constituée de n_h (sur N_h) UPB dans chaque strate H . Les USE dans les UPB échantillonnées sont ensuite stratifiées à leur tour en D domaines. Dans chaque domaine d , on prélève un sous-échantillon de m_d (sur M_d) USE. Aux deux phases du plan de sondage, on a procédé à un éas stratifié sans remise.

¹ Phillip S. Kott, statisticien principal, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, D.C. 20250, USA.

- CASSEL, C.M. (1978). Probability based disclosure. Dans Dalenius, T., et Klevmarcken, A. (éds.), *Proceedings of a symposium on personal integrity and the need for data in the social sciences*. Swedish Council for Social Science Research, Stockholm, 189-193.
- CASSEL, C.M. (1978). On errors in the predictions with logit models. Rapport technique, Statistics Sweden.
- DALENIUS, T. (1953). Något om metoder för objektiva skördeberäkningar. (About methods for objective crop estimation.) *Kungliga Lantbruksakademiens Tidskrift*, 92, 99-118.
- DALENIUS, T. (1957). *Sampling in Sweden. Contributions to the Methods and Theory of Sample Survey Practice*. Stockholm: Almqvist and Wiksell.
- DALENIUS, T. (1988). Controlling Invasion of Privacy in Surveys. Statistics Sweden.
- INTERNATIONAL STATISTICAL INSTITUTE (1986). Declaration of Professional Ethics. *International Statistical Review*, 54, 227-242.
- LUNDSTRÖM, S. (1987). Utveckling av estimatorer för skatning av antal förärvsarbete i olika arbetsstidsklasser inom små redovisningsgrupper. R&D Report, U/STM 40, Statistics Sweden.
- LYBERG, L. (1981). Control of the coding operation in statistical investigations. Uval n° 13, Statistics Sweden.
- STATISTICS SWEDEN (1987). Statistics and Privacy: Future Access to Data for Official Statistics – Cooperation or Distrust? Statistics Sweden.
- SWENSSON, B. (1977). Survey measurement of sensitive attributes. Thèse de Ph.D., Département de statistique, Université de Stockholm.

12. Modélisation combinée avec les principes classiques de l'échantillonnage au hasard. Depuis les années 50, les méthodes qui s'appliquent aux enquêtes ont suivi de près la tradition bien établie, en matière d'échantillonnage au hasard, due à Neyman ainsi qu'à Hansen et à ses collègues aux Etats-Unis. Cependant, il faut parfois avoir recours à la modélisation dans des enquêtes quand la théorie classique de l'échantillonnage au hasard ne suffit pas. Depuis les années 70, on a étudié sous tous ses aspects l'utilisation de la modélisation dans les enquêtes. Le livre *Foundations of Inference in Survey Sampling* de Cassel, Särndal et Wretman (1977) a exposé les nouvelles tendances. De plus, un certain nombre d'articles par ces auteurs et par d'autres Suédois ont montré comment des modèles peuvent être utiles à l'inférence à partir d'enquêtes. Au cours des dernières années, les méthodologistes du service de la statistique de la Suède ont fait preuve d'une ouverture d'esprit inhabituelle pour ce qui est de se servir de modèles afin d'établir des estimations faites à partir d'enquêtes. "L'enquête d'Oresund", qui sert à mesurer le débit de circulation entre la Suède et le Danemark, constitue un des premiers exemples d'une enquête où l'on a combiné des idées fondées sur un plan avec des idées fondées sur un modèle. Le plan de sondage est étudié dans Cassel (1978). Certaines enquêtes sont maintenant conçues à l'aide d'hypothèses de modélisation, comme c'est le cas pour l'enquête sur la population active décrite dans Lundström (1987) et dans un projet en cours visant à restructurer le secteur des enquêtes-entreprises.

13. Protection des renseignements personnels dans les enquêtes. Au cours des deux dernières décennies, le grand public s'est de plus en plus préoccupé des atteintes à la vie privée relativement aux enquêtes réalisées par le service de la statistique de la Suède, y compris les recensements de la population. Par suite de cette situation, les taux de non-réponse ont eu tendance à croître pour certaines enquêtes. Plusieurs mesures ont été prises pour régler ce problème: i) Le service de la statistique de la Suède a adopté la déclaration d'éthique de l'Institut international de statistique (1986); une traduction de cette déclaration a été distribuée à tous les employés; ii) en 1987, le service de la statistique de la Suède a été l'hôte d'une conférence internationale qui s'est concentrée sur des questions de politique (pour les distinguer des "techniques"); les discussions qui ont eu lieu dans le cadre de cette conférence sont résumées dans *Statistics Sweden* (1987); iii) le service de la statistique de la Suède a favorisé l'élaboration de nouvelles mesures de protection des renseignements personnels dans ses enquêtes et il a pris des mesures actives pour les appliquer. Une étude de ces mesures est donnée dans Dalenius (1988). L'article de Block et Olsson (1976), qui décrit une mesure du pouvoir d'identification des quasi-identificateurs et celui de Cassel (1976), qui porte sur la divulgation basée sur les probabilités présentent un intérêt exceptionnel.

14. Evénements particuliers. L'appréciation croissante des enquêtes par sondage depuis 1950 environ a amené la création, en 1953, du centre de recherches sur les enquêtes du service de la statistique de la Suède. On pourrait donner une interprétation semblable à l'établissement d'un poste de professeur en "statistiques, particulièrement en statistiques officielles" à l'Université de Stockholm en 1965. De plus, des postes de professeurs en méthodologie d'enquête ont été créés récemment au service de la statistique de la Suède.

BIBLIOGRAPHIE

BLOCK, H., et OLSSON, L. (1976). Bakvägsidentifiering. (Backwards identification) *Statistisk Tidskrift*, 1976, 135-144.

CASSEL, C.M., SÄRNDAL, C.E., et WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.

sélection, relativement parlant, d'un plus grand nombre de grosses fermes que de petites fermes. – Il est intéressant de remarquer que les autorités gouvernementales responsables de l'évaluation du plan de sondage ont jugé nécessaire de consulter la sous-commission des N.U. sur l'échantillonnage statistique à propos de l'opportunité de cette déviation radicale par rapport à la "règle de toutes les dixièmes unités". La sous-commission a approuvé de tout coeur le plan de sondage. Il a donc été adopté en principe. Le plan de sondage a offert beaucoup d'occasions d'effectuer des recherches. En fait, trois contributions à la théorie de l'échantillonnage stratifié en ont découlé, notamment : i) comment diviser le mieux possible une population en L strates; ii) comment choisir le mieux possible le nombre de strates et iii) comment répartir l'échantillon entre les strates pour estimer plusieurs paramètres. Le plan de sondage proposé prévoyait aussi que l'on devait s'attaquer au problème des "erreurs de mesure" relatives à la superficie et l'on proposait la réalisation d'une enquête de calibration spéciale. Cependant, les autorités ont rejeté cette proposition.

9. Estimation du rendement. Au cours de la Seconde Guerre mondiale, le rendement de diverses cultures a été estimé à l'aide de données recueillies par des estimations au jugé du rendement par unité de surface. En 1950, on avait réalisé que cette méthode de collecte des données pouvait être sérieusement biaisée. Au début des années 50, le temps était mûr pour considérer une approche différente, nommément, l'estimation des cultures basées sur la récolte faite sur une placette d'échantillonnage, ce que l'on appelle "l'estimation objective des cultures". Une étude pilote a donc été réalisée pour faire l'essai de l'utilisation de cette approche. Le résultat de l'essai a été convaincant. Depuis ce temps, on a utilisé la méthode "objective". Dans le cadre de la conception de l'enquête pilote, on a élaboré une méthode pour effectuer la sélection sans remise de $n = 2$ fermes tirées d'une strate avec probabilité proportionnelle à la taille, comme cela est décrit dans Dalenius (1953). Selon cette méthode, il fallait diviser chaque strate au hasard en deux parties puis choisir une ferme dans chacune de ces parties.

10. Progrès relatifs aux erreurs non dues à l'échantillonnage. Au début des années 50, en Suède comme dans les autres pays, on a accordé une attention considérable au problème de la non-réponse. Il n'était pas rare que l'on ait des enquêtes avec des taux de non-réponse de 20 à 30%. Cela a donné lieu à une discussion vive et parfois passionnée dans la communauté des statisticiens à propos de la distorsion possible des estimations. Pendant un certain temps, il a semblé que les statisticiens avaient maîtrisé le problème. Dernièrement, les inquiétudes du public à propos des atteintes à la vie privée ont changé cette situation; la non-réponse est redevenue un problème sérieux. – Au cours des 15 dernières années, plusieurs contributions ont été faites dans le domaine de la limitation de l'erreur non due à l'échantillonnage. Le problème du "biais dû aux réponses évasives", pour employer le terme utilisé pour la première fois par S. Warner en rapport avec la réponse randomisée a été abordé dans Swensson (1976). Et Lyberg (1981) s'est attaqué avec succès au problème du contrôle de l'opération de codage dans un recensement de la population ou dans une enquête avec interviews.

11. Fardeau des répondants. Au cours des dernières années, on s'est préoccupé de plus en plus du fardeau des répondants et de ses effets négatifs sur les taux de réponse. Par exemple, la population cible dans de nombreuses enquêtes-entreprises est la même population plutôt limitée. Le problème peut être atténué par l'emploi de techniques spéciales de sélection de l'échantillon. Le système SAMU utilisé pour les enquêtes-entreprises par le service de la statistique de la Suède permet le "coordonation négative" des échantillons, dans le sens où des échantillons sans chevauchement peuvent être choisis à l'aide de la technique connue sous le nom de JALRS. On joint à chaque unité dans la base de sondage un nombre aléatoire distribué uniformément. Ce nombre reste avec l'unité et il est utilisé pour sélectionner des échantillons dans le temps.

4. L'enquête de 1911 sur les forêts dans la province de Värmland. La caractéristique essentielle du plan de sondage était le fait que le volume de bois était mesuré sur des placettes d'échantillonnage le long de bandes de 10 mètres de largeur qui couvraient cette province. Il vaut la peine de remarquer que l'on a fait appel à la théorie de la probabilité pour analyser les "caractéristiques représentatives" de l'enquête.

5. L'enquête de 1911 sur le logement à Göteborg. Cette enquête a été réalisée par le bureau de la statistique de la ville de Göteborg. La sélection d'un échantillon d'appartements était basée sur un schéma d'urnes. Chaque immeuble de Göteborg était représenté par un bout de papier sur lequel figuraient des renseignements permettant de l'identifier. Les bouts de papier ont été bien mélangés dans une urne et on a choisi un échantillon de 20% des bouts de papier. Le plan de sondage visait à empêcher que l'on puisse dire que l'enquête utilisait un échantillon biaisé. Le schéma d'urnes était décrit par la personne responsable de l'enquête comme la seule méthode "que l'on peut dire représentative".

6. Le recensement partiel de la population de 1935/1936. Cette enquête par sondage utilisait un plan de sondage élaboré prévoyant une sélection contrôlée. Les résultats de ce recensement ont joué un rôle décisif dans un débat intense, en Suède, sur une "crise de la population" que l'on craignait à la suite des faibles taux de fécondité à ce moment.

III. LA PÉRIODE POSTÉRIEURE À 1950

7. Les débuts d'une nouvelle ère. Les communications internationales grandement améliorées après la fin de la Seconde Guerre mondiale ont contribué à mettre les statisticiens suédois au courant des progrès réalisés depuis peu dans la théorie, les méthodes et les applications des enquêtes par sondage aux États-Unis et en Inde, pour mentionner deux des pays à l'avant-garde dans ce domaine. Les nouveaux progrès ont été étudiés et ils ont fait l'objet de discussions, par exemple, lors de la conférence des statisticiens scandinaves tenue à Helsinki en 1949. Les statisticiens étaient fiers de pouvoir employer le "vocabulaire des méthodes utilisées pour les enquêtes par sondage"; il est certain, que dans certains cas, cette compétence se limitait à la connaissance de certains termes techniques, notamment le mot "stratification". - Il faudrait aussi mentionner l'influence exercée par les Nations Unies et ses organismes affiliés comme l'Organisation des Nations Unies pour l'alimentation et l'agriculture. - Dans les paragraphes ci-après, nous donnons des exemples d'enquêtes par sondage et de progrès connexes dans les méthodes et la théorie. Pour les cas remon-

8. L'inventaire par sondage des superficies et des bestiaux de 1950. Au cours des années 30, des enquêtes par sondage ont été utilisées pour estimer la superficie de diverses cultures ainsi que le nombre de bestiaux. On désignait ces enquêtes par l'expression "comptages représentatifs". Elles étaient basées sur une sélection non probabiliste des fermes. Le but visé, qui n'a pas été atteint cependant, était de choisir 1/10 des fermes dans chacune de plusieurs tranches de taille entre lesquelles les fermes avaient été réparties. Au cours des années 40, ces enquêtes ont été réalisées par dénombrement complet. Pour l'enquête de 1950 on a décidé d'utiliser à nouveau l'échantillonnage. Le plan d'échantillonnage qui a été proposé et appliqué dans une large mesure pour cette enquête représentait une rupture partielle par rapport à la tradition classique qui consistait à choisir toutes les dixièmes unités. Bien que la taille totale de l'échantillon ait été fixée par les autorités gouvernementales à 1/10 du nombre total de fermes dans la population cible, le nouveau plan de sondage prévoyait la stratification des fermes selon des tranches de taille basées sur la superficie et l'emploi d'une répartition visant à obtenir une variance minimum, ce qui supposait la

Certains progrès relatifs aux techniques de sondage et à leur utilisation dans les statistiques officielles en Suède

TORÉ DALENIUS et CARL-ERIK SÄRNDAL¹

Dans le présent article nous présentons certaines caractéristiques importantes de l'histoire des enquêtes par sondage en Suède et nous commentons les progrès connexes dans les techniques de sondage (méthodes et théorie) utilisées dans les statistiques officielles. Le compte rendu est divisé en trois périodes, de la façon suivante: i) avant 1900; ii) de 1900 à 1950 et iii) après 1950. L'accent est mis sur la troisième période.

I. LA PÉRIODE AVANT 1900

1. Un résumé. Comme cela est décrit dans Dalenius (1957), il y a eu une résistance évidente contre les enquêtes par sondage dans les domaines traditionnels des statistiques officielles, les enquêtes par sondage étaient justifiées principalement dans les cas où les circonstances ne permettaient pas de réaliser des recensements. Dans d'autres domaines il y avait, cependant, des signes d'appréciation, comme on le voit dans la section suivante.

2. Deux exemples classiques. Au cours des années 1820, la superficie des prés en Suède était estimée à l'aide de la technique suivante. Pour chaque comté pris séparément, on calculait le rapport entre la superficie en terres arables pour un échantillon de fermes. Ce rapport était alors appliqué à la superficie totale des terres arables du comté pour laquelle une estimation distincte était disponible. Et, en 1830, un administrateur d'une commission des forêts a proposé d'estimer le volume de bois dans une forêt à l'aide d'un "inventaire par échantillonnage en bandes".

II. LA PÉRIODE DE 1900 À 1950

3. Les principales caractéristiques. Les possibilités qu'offraient les enquêtes par sondage en matière de statistiques officielles commençaient lentement à être comprises. Dans la mesure où des enquêtes par sondage ont été utilisées pendant cette période, le plan de sondage prévoyait généralement un échantillonnage systématique, chaque fois que cela était réalisable en pratique. Dans de nombreuses applications, la fraction de sondage était de 1/10 ou de 1/5. – Dans les années 40, l'économie de guerre, avec ses règlements et son rationnement, constituait un facteur qui favorisait les recensements. Cette influence qui a duré à peu près jusqu'à la fin de cette décennie, était cependant contrebalancée par l'introduction des sondages Gallup en Suède et, plus particulièrement, par la précision impressionnante des prévisions de l'Institut Gallup pour les élections de 1944. En particulier, ces tendances étaient suivies avec intérêt par les personnes chargées de recueillir les statistiques officielles.

¹ Toré Dalenius, Brown University, Carl-Erik Särndal, Université de Montréal.
Les circonstances ont empêché les auteurs de discuter du contenu de cet article avec des représentants du service de la statistique de la Suède.

Scheuren est très flatté quand il dit que les recensements par étapes constituent un nouveau paradigme. Il est vrai que, comme tous les nouveaux paradigmes, quand je les présente, les agrégations et les échantillons avec renouvellement complet, font face à trois gros blocages psychologiques: a) la mise en moyenne de données obtenues à des dates variables plutôt qu'à une date arbitraire comme le 1er avril de l'année du recensement décennal; b) l'acceptation d'une certaine mobilité des populations humaines plutôt que de fixer tout le monde à des endroits uniques; c) l'emploi d'échantillons avec renouvellement complet pour remplacer des secteurs-échantillons primaires fixes. Il peut donc sembler paradoxal que Morris Hansen remarque que mon analyse de ces formes d'échantillon "n'apporte rien de nouveau." Il est possible que Hansen ait déjà rencontré toutes ces propositions et qu'il en ait peut-être écarté un certain nombre. Personnellement, j'ai décrit les échantillons avec renouvellement complet depuis au moins 1961 et proposé les recensements par étapes depuis 1965. Mais j'ai aussi constaté que, pour de nombreuses personnes, il s'agit de nouvelles idées et souvent de nouvelles idées étranges.

Finalement, permettez-moi d'ajouter deux renseignements importants sur les origines des sondages dans les années 40, bien que, pour moi, les personnes et les priorités ne soient que des aspects mineurs de l'histoire de toute science. Il faudrait mentionner l'université Iowa State à Ames, où, sous la direction de George Snedecor et de Henry Wallace, Bill Cochran a donné, au printemps 1939, le premier cours d'échantillonnage et d'où sont sortis les premiers détenteurs de maîtrise puis de doctorat en échantillonnage. Ensuite, Henry Wallace (à nouveau) a été à l'origine de la Division of Program Surveys du Département de l'Agriculture des E.-U.; il m'a embauché en 1941, puis il a engagé Steve Stock en 1942 pour les premiers échantillons nationaux à Washington au cours de cette dernière année, vint ensuite l'échantillon de 1943 au USB-C. Stock, Frankel et Webb (de la WPA (Work Projects Administration)) ont commencé à donner le deuxième cours d'échantillonnage, à l'automne 1939, à l'école d'études supérieures du Département de l'Agriculture (USDA) qui, sous la direction de Hansen, de Hurwitz et de leurs collègues du Bureau of the Census, est devenue célèbre et productive. Parmi les cours importants, donnés dans cette école, je témoignerai particulièrement de ceux de Deming, la figure dominante de l'école. Au cours des années quarante, l'enseignement ainsi que l'étude de l'échantillonnage se sont surtout faits à Ames et au USDA ainsi qu'au USB-C.

Réponse

LESLIE KISH

Dans son excellente analyse Fritz Schuren complète nos comparaisons des méthodes qui peuvent remplacer les recensements en préconisant l'emploi de fichiers administratifs pour les E.-U. J'appuie son plaidoyer expert pour étudier ce que ces méthodes pourraient offrir comme ajoutés dans de nombreux pays et nous aimerions savoir où, quand et comment? Il est même utilisés dans de nombreux pays et nous aimerions savoir où, quand et comment? Il est même probable que, bientôt, ils ne serviront pas seulement à compléter les recensements décennaux dans certains pays, ils les remplaceront. Quand cela se produira-t-il aux E.-U.? Je ne le sais pas; ce n'est qu'assez tardivement et lentement que notre pays a adopté un registre des naissances et des décès qui existe encore. Et même actuellement la déclaration de ces événements se fait plutôt lentement.

Les échantillons avec renouvellement complet pourraient être conçus de façon à ce que la déclaration se fasse rapidement et l'actualité des données n'est qu'un des avantages qu'offrent ces échantillons. C'est donc faire preuve d'un manque d'objectivité que de comparer les recensements par étapes avec les recensements traditionnels, tant pour ce qui est des coûts que des avantages, en ne se basant que sur le seul produit pour l'obtention duquel les recensements décennaux ont été conçus. Il faudrait des études techniques détaillées pour comparer les facteurs relatifs aux coûts, à la couverture, à l'actualité, au contenu, etc. des recensements par étapes par opposition aux recensements décennaux aux E.-U. Mais, on peut réaliser beaucoup de choses avec entre 10 et 15 millions de dollars par mois. La question des recensements adéquats est très saillante en 1990 aux E.-U. et ailleurs, mais on ne devrait pas oublier les autres utilisations des échantillons au moment où nous effectuons la planification pour la dernière décennie du vingtième siècle.

Ma contribution vise surtout à développer l'intérêt relatif aux divers avantages des agrégations faites à partir d'échantillons périodiques, qui ont été négligés en faveur d'autres avantages que l'on peut tirer d'un nombre croissant d'enquêtes périodiques. Il se peut qu'un jour les recensements par étapes deviennent un de ces avantages et les échantillons avec renouvellement complet ont déjà été utilisés – bien que je pense qu'ils ne l'aient pas été assez souvent. Il existe très peu d'agrégations asymétriques et elles sont obscures et les plans à panel fractionné que je propose sont complètement absents.

De plus, je n'ai pas seulement des visées nationales (les E.-U.), pas même continentales (l'Amérique du Nord); elles sont intercontinentales et internationales. On a, par exemple, commencé à utiliser les fichiers administratifs dans les pays scandinaves et il se peut qu'on les emploie au Canada avant de le faire aux E.-U. L'emploi des recensements par étapes implique un accroissement beaucoup moins considérable des enquêtes sur la population active au Canada parce que la taille de la population n'est que le dixième de celle des E.-U., comme Fritz et moi le montrons. Mais il se peut bien que d'autres pays utilisent les recensements par étapes avant soit le Canada, soit les Etats-Unis.

Les échantillons avec renouvellement complet et les agrégations ne peuvent pas seulement être employés à l'échelle internationale, ils peuvent aussi être interdisciplinaires, ils ne sont pas réservés à la seule production de chiffres de population. Un bon nombre des autres besoins des bureaux statistiques – et des autres organismes qui recueillent des statistiques – seraient mieux servis par un personnel "permanent" bien entraîné que par une grosse armée embauchée à la hâte et dont le temps de formation est à peu près égal à la brève période d'emploi.

Réponse

BARBARA BAILLAR

Les commentateurs des participants décrivent encore plus les contributions de l'administration fédérale au domaine de la statistique. Je tiens remercier Gordon Brackstone et Morris Hansen d'avoir mentionné ces sujets supplémentaires. Le sujet que j'ai oublié et qui pourrait avoir l'incidence la plus forte sur la statistique et les autres domaines quantitatifs fait l'élaboration de l'ordinateur pour le traitement et l'analyse des données. Là encore, l'équipe Hansen-Hurwitz était l'avant-garde, pressant et finançant l'élaboration de l'UNIVAC I et ensuite l'installant au Census Bureau pour aider au traitement du recensement de 1950.

Morris Hansen a présenté la remarquable équipe au Census qui a travaillé avec lui et Bill Hurwitz sur tant de domaines. Je crois que j'ai eu beaucoup de chance d'avoir pu commencer ma carrière au Census Bureau lorsque ces personnes étaient là et d'avoir pu travailler avec la plupart d'entre elles pendant de nombreuses années. Il est rare que l'on puisse bénéficier d'un tel apprentissage.

Gordon Brackstone demande si la méthodologie statistique mise au point par le Census Bureau présentait des avantages pour l'ensemble de la statistique. Certes, compte tenu de l'intégration des bureaux de statistiques nationales, le Bureau of the Census a influencé les activités statistiques gouvernementales dans les autres pays. Brackstone estime que l'incidence de l'élaboration du Census Bureau sur les départements de statistiques des universités a eu des résultats mitigés. Il a peut-être raison en ce qui concerne les cours, mais je pense que le programme de bourses ASA-NSF-Census et le programme Agriculture Fellowship ont eu une incidence plus grande. Davantage de professeurs et d'étudiants universitaires de deuxième cycle sont au courant des problèmes des erreurs non dues à l'échantillonnage, de la non-divulgateion et des séries chronologiques et y travaillent. La mise en place récente de programmes de bourses au Bureau of Labor Statistics et au National Center for Education Statistics également mis en relief ces recherches entreprises dans l'un des organismes gouvernementaux.

Le problème principal maintenant est l'application des résultats des recherches. Beaucoup de programmes gouvernementaux ne mettent que lentement en oeuvre les méthodologies nouvelles parce que le changement est perturbateur. Mais le changement est nécessaire si l'on veut améliorer les méthodes.

Nous nous demandons toutefois si les progrès technologiques qui sont signales de part et d'autre, conjugués à la multiplicité des enquêtes qui est aussi signalée de part et d'autre, n'auraient pas autant de conséquences négatives que positives. Par exemple, lorsque des étudiants du premier cycle universitaire (ou n'importe quel débutant dans le domaine de l'analyse statistique) font des analyses complexes de données d'enquête à l'aide de logiciels statistiques, il arrive souvent qu'ils ne comprennent pas les données analysées et que les méthodes statistiques qu'ils emploient ne soient pas appropriées.

La multiplicité des enquêtes tient non seulement à la demande accrue de renseignements mais aussi à la facilité relative avec laquelle on peut réaliser des enquêtes et analyser des données dans l'état actuel de la technologie. (Nous croyons en outre qu'il sera de plus en plus facile d'avoir accès à des données d'enquête pour réanalyse grâce aux nouvelles techniques de stockage comme le disque CD-ROM et le disque optique.) Toutefois, une telle facilité est un avantage incertain. Comme l'indique Groves, le phénomène de non-réponse a pris de l'ampleur aux Etats-Unis dans les années 1980; ce phénomène s'était manifesté plus tôt et dans une plus forte mesure (du moins jusqu'à maintenant) en Europe. Il n'est pas nécessaire de postuler un plus grand besoin de protection de la vie privée pour expliquer la baisse des taux de réponse, bien qu'un tel besoin puisse réellement exister. Il suffit de voir les problèmes de non-réponse qu'éprouve le U.S. Bureau of the Census dans le recensement décennal de 1990, qu'il s'agisse des questionnaires retournés par la poste ou des interviews sur place, pour se convaincre du fait que les répondants commencent à en avoir assez des enquêtes.

Par ailleurs, comme le souligne Groves, la théorie et la pratique des sondages n'ont pas été l'apanage des universitaires puisque les sondages n'ont pas constitué avec le temps une discipline propre ayant des normes et des critères de formation précis. Comme il n'existe pas de département de sondages dans les universités, à peu près n'importe qui moiindrement intéressé peut réaliser sa propre enquête ou analyser des données d'enquête. Tandis que certaines personnes exécutent ces tâches convenablement, d'autres les exécutent pitoyablement, ce qui a pour effet de discréditer tout le processus d'enquête. Par conséquent, si nous devons présenter en l'an 2040 le rapport optimiste que Groves envisage, nous croyons qu'il sera nécessaire d'institutionnaliser les innovations qui pourront surgir entre temps au chapitre de l'enseignement et de la formation.

Nous sommes particulièrement heureux de voir M. Hansen rehausser par ses commentaires notre bref compte rendu de l'évolution des sondages au sein de l'administration publique des E.-U. dans les années 1930 et 1940. Grâce à ses commentaires, M. Hansen ajoute à notre article la dimension humaine dont Groves déplore l'absence au départ.

Hansen étotte aussi nos propos concernant le lien entre les recensements et les sondages et l'introduction de l'autodénombrement dans le recensement des Etats-Unis, qui a été décidée par suite d'une étude des erreurs de réponse. Compte tenu de la forte hausse du taux de non-réponse au chapitre de l'autodénombrement dans le recensement décennal de 1990, il faudrait peut-être revoir l'incidence des diverses composantes du modèle de Hansen-Hurwitz-Marks-Mauldin pour les erreurs non dues à l'échantillonnage. Par ailleurs, souligons que, dans le cadre du recensement de 1990, le U.S. Bureau of the Census créera une nouvelle enquête post-censitaire (EP) qui couvrira 150,000 ménages et dont les résultats serviront à évaluer le taux de sous-dénombrement dans le recensement. Les progrès techniques réalisés dans la gestion informatique des données et l'appariement informatique de fichiers de l'EP et du recensement ont rendu possible cette nouvelle enquête et l'utilisation de ses résultats comme mesures du taux de sous-dénombrement et de surdénombrement des ménages.

SOURCES ADDITIONNELLES

Turner, C.F. et Martin, E. (éd.) (1984). *Surveying Subjective Phenomena*. Vol. 1. New York: Russell Sage Foundation.

Réponse

STEPHEN E. FIENBERG et JUDITH M. TANUR

Nous sommes reconnaissants envers Robert Groves et Morris Hansen pour leurs commentaires perspicaces et envers le rédacteur en chef de *Techniques d'enquête*, qui nous donne l'occasion de nous exprimer des maintenant plutôt qu'en 2040. Messieurs Groves et Hansen soulèvent plusieurs points importants; nous allons aborder ces points un à un.

Nous sommes entièrement d'accord avec Groves lorsqu'il affirme que les administrations publiques qui mettent l'accent sur les programmes sociaux recherchent plus souvent des renseignements que celles qui poursuivent d'autres buts. La preuve que les propos de Groves sont fondés est le fait que l'enquête nationale la plus importante qui ait été mise sur pied aux États-Unis dans les années 1980, années qui n'ont pas été particulièrement propices à l'expansion des programmes sociaux aux États-Unis, est la Survey of Income and Program Participation, qui a notamment pour objectif d'évaluer l'effet des programmes sociaux de l'administration américaine sur le revenu et l'actif. De plus, une autre illustration de la justesse des propos de Groves est la tendance des pays d'Europe de l'Est, dans la vague de démocratisation, à demander l'aide des pays occidentaux pour améliorer leurs systèmes statistiques et notamment mettre sur pied des infrastructures pour la réalisation de grandes enquêtes. Groves semble donc appuyer notre affirmation selon laquelle les bases institutionnelles des enquêtes par sondage déterminent le contenu et l'orientation de telles enquêtes. La question de savoir si ces bases institutionnelles ont été le foyer d'une élite en cette matière nous semble plutôt théorique – cette interrogation doit être vue plus comme un cadre d'analyse que comme une question qui exige une réponse catégorique. Nous sommes bien sûr d'accord avec Groves lorsqu'il dit que les objectifs des divers groupes d'institutions ont dicté, du moins en partie, le choix des activités auxquelles se livrent ces institutions. Toutefois, à l'instar de Groves, qui souligne l'urgence d'adopter une perspective transnationale, nous constatons que les rôles des institutions varient selon les pays. Par exemple, beaucoup sont d'avis aux États-Unis que l'administration fédérale ne devrait pas s'occuper de recueillir des données d'enquête sur des phénomènes subjectifs (voir par exemple Turner et Martin 1984, p. 31-39) – pour sa part, le gouvernement britannique ne voit pas les choses de la même façon, à en juger surtout par son rapport annuel *Social Trends* (Turner et Martin 1984, p. 4).

Groves laisse à entendre que la démarcation entre les divers groupes d'institutions (universités, maisons de sondage, administrations publiques) n'est pas aussi perméable que nous voulons le laisser croire. Ni lui ni nous ne disposons de données empiriques sur la question, mais nous tenons ici à rappeler l'importance des institutions intermédiaires, qui réunissent des représentants des divers secteurs et servent de lieu d'échange sur le plan professionnel sinon sur le plan personnel, et nous nous empressons d'ajouter que la nomination récente de Groves au poste de directeur associé du U.S. Bureau of the Census de même que le cheminement inverse suivi par Hansen il y a une vingtaine d'années illustrent bien l'interpénétration des secteurs à défaut d'en illustrer le degré.

Groves suppose indirectement que nous avons choisi de nous concentrer sur les développements technologiques, les enquêtes longitudinales et les aspects cognitifs des enquêtes parce que ce sont là les domaines pour lesquels nous manifestons le plus d'intérêt et dans lesquels nous avons le plus d'expérience, et il signale plusieurs autres marques de progrès qui méritent d'être considérées. Il a raison de supposer que nous sommes arrêtés aux questions qui nous intéressent le plus; seulement, la question des progrès technologiques intègre sans aucun doute les deux premiers sujets d'intérêt additionnels proposés par Groves, à savoir l'élaboration de logiciels statistiques généraux et l'existence de fichiers permanents de données d'enquête.

Nous sommes aussi convaincus, comme Smith, de l'importance de mesurer l'erreur totale des enquêtes permanentes.

SOURCES ADDITIONNELLES

- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society, Indian Statistical Institute.
- MAHALANOBIS, P.C. (1938). *Statistical report on the experimental crop census*, 1937. Indian Central Jute Committee.
- RAO, J.N.K., et GHANGURDE, P.D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association*, 67, 439-443.
- SHAO, J., et WU, C.F.J. (1989). A general theory for jackknife variance estimations. *Annals of Statistics*, 17, 1176-1197.

Réponse

J.N.K. RAO et D.R. BELLHOUSE

Nous sommes reconnaissants envers messieurs Hansen et Smith pour leurs commentaires utiles.

Morris Hansen a fait des observations importantes sur le développement de l'échantillonnage avec PPT. Il a raison de dire que Hansen et Hurwitz (1943) n'ont pas proposé l'utilisation de l'échantillonnage avec remise et qu'ils n'ont supposé ce genre d'échantillonnage que pour l'estimation de la variance. À propos, Murthy (1967, p. 184) signale que Mahalanobis (1938) s'est intéressé à l'échantillonnage avec PPT ainsi qu'à l'estimateur non biaisé correspondant d'un total pour ce qui a trait à l'échantillonnage de parcelles dans les enquêtes sur les cultures.

Hansen a fait aussi des observations intéressantes sur l'utilisation de l'estimateur «jackknife» pour des fonctions non lisses comme les quantiles. Il ne fait plus de doute que, en ce qui a trait aux quantiles, cet estimateur est inconsistant selon un échantillonnage aléatoire simple. Des résultats empiriques consignés par Kovar, Rao et Wu (1988) montrent qu'il est aussi inconsistant tant selon un échantillonnage aléatoire simple stratifié. Il est aussi susceptible d'être inconsistant selon un échantillonnage en grappes stratifié si les sous-échantillons tirés des grappes sont petits ou que les corrélations intra-grappe sont appréciables. Dans le cas de l'estimateur utilisé par Hansen, les sous-échantillons sont assez grands et les corrélations intra-grappe, minimes. En l'occurrence, il est permis de croire que l'estimateur de variance «jackknife» utilisé par Hansen est conforme à la conclusion de Shao et Wu (1989) selon laquelle l'estimateur «jackknife» avec suppression de d éléments est convergent selon un échantillonnage aléatoire simple pourvu que $n^{1/2}/d \rightarrow 0$ et $n-d \rightarrow \infty$ lorsque la taille d'échantillon $n \rightarrow \infty$.

La méthode proposée par Hansen et qui consiste à diviser un échantillon aléatoire simple en m sous-échantillons de taille d par exemple puis à supprimer un sous-échantillon à la fois, se rapproche sensiblement de la méthode «jackknife» de Shao et Wu sauf que ceux-ci considèrent tous les sous-échantillons possibles ($\frac{n}{d}$) dans la construction de l'estimateur de variance. Par ailleurs, l'estimateur «jackknife» de Shao et Wu risque d'être plus stable. Shao et Wu s'intéressent aussi au sous-échantillonnage équilibré qui n'exige que b sous-ensembles de taille $n-d$, où $b \geq n$ est le nombre de blocs dans un plan en blocs équilibrés incomplets. Smith a fait des observations importantes sur les aspects fondamentaux de la théorie des sondages et a fait ressortir l'intérêt de l'étude réalisée par Ericson (1969) sur l'estimation bayésienne de totaux suivant des distributions a priori interchangeables. Soulignons à ce propos que Hartley et Rao (1968) avaient obtenu des résultats équivalents à ceux de Ericson pour la moyenne et la variance a posteriori suivant un échantillonnage aléatoire simple. Dans son analyse de l'article d'Ericson, A. Scott n'a pas manqué de souligner cette similitude. Toutefois, un avantage de la méthode de Hartley-Rao par rapport à celle d'Ericson est que l'inférence repose sur le plan de sondage. De plus, la méthode de Hartley-Rao permet de faire des inférences classiques utiles. Rao et Changurde (1972) ont appliqué les résultats de Hartley et Rao à l'échantillonnage aléatoire stratifié, à l'échantillonnage double avec taille de strate inconnue, à la méthode de traitement de la non-réponse de Hansen et Hurwitz ainsi qu'à l'échantillonnage aléatoire à deux degrés.

Le modèle global (TV - théorie universelle) que propose Smith pour faire des inférences offre de très belles perspectives. Nous sommes d'accord avec ce participant lorsqu'il affirme qu'il y a peu de différence dans la pratique entre les estimateurs ponctuels de l'une ou l'autre des méthodes proposées et que le problème consiste essentiellement à choisir une mesure de variabilité, comme nous l'indiquons d'ailleurs dans notre article.

Je suis de ceux qui croient qu'un recensement quinquennal et les grandes enquêtes permanentes qui existent actuellement valent bien les sommes considérables qu'on leur consacre. Cependant, si nous devions recourir au recensement par étapes, comme le propose Kish, j'estime qu'il faudrait utiliser des échantillons chevauchants. Un recensement par étapes, même sans échantillon chevauchant, pourrait revenir beaucoup plus cher que le programme de recensement actuel qui comprendrait aussi un recensement quinquennal. Il faut se demander si un recensement par étapes vaut les dépenses additionnelles qu'il nécessiterait ou s'il présente des avantages notables par rapport à un recensement quinquennal et à des enquêtes intercensitaires d'envergure. Je crois que, dans la plupart des cas, un recensement par étapes produirait des données moins utiles que celles d'un recensement quinquennal puisqu'il s'agirait uniquement de valeurs moyennes pour une période de 10 ans. Du point de vue de l'analyse coûts-avantages, il semble plus avantageux d'opter pour un recensement quinquennal, accompagné d'enquêtes à passages répétés (pour obtenir des données relativement récentes sur les grandes régions) et d'autres types de sondages visant à produire des données sur les États et les comtés et peut-être aussi des estimations démographiques pour les petites régions.

Kish mérite notre estime pour avoir proposé des solutions radicalement nouvelles à quelques-uns des problèmes que pose le recensement classique. Toutefois, à mon humble avis, le recensement par étapes ne semble pas être la solution recherchée. Peut-être qu'une utilisation plus efficace des fichiers administratifs donnerait des résultats plus encourageants, en supposant toujours que cette option soit complétée par des enquêtes permanentes et un recensement décennal ou même, souhaitons-le, quinquennal. Il y a peut-être lieu de fonder beaucoup d'espoir dans le nouveau système de cartographie et de codage informatisé mis au point par le Cens Bureau pour le recensement de 1990; ce système, qui est connu sous le nom de TIGER, pourrait bien améliorer les méthodes du recensement et des enquêtes actuelles. De plus, l'intégration de ce système aux principaux fichiers administratifs pourrait accroître les possibilités d'utilisation de ces fichiers, par exemple pour l'établissement d'estimations de population. Il est à souhaiter que le Cens Bureau crée bientôt un programme de mise à jour du système TIGER ainsi qu'un registre d'adresses constamment à jour.

BIBLIOGRAPHIE

- HANSEN, M., HURWITZ, W., et MADOW, W. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.
- HANSEN, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2, 180-190.
- OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 162-179.
- DUNCAN, J.W., et SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. U.S. Government Printing Office, Washington, D.C.

présomme qu'il les présente dans un contexte très général et qu'il cherche à tracer un lien entre ces échantillons et le recensement par étapes.

Le recensement par étapes dont il est question dans l'article de Kish est une enquête hebdomadaire où la population totale des unités de logement est divisée en 520 sous-échantillons non chevauchants; le recensement de la population se fait au rythme d'un sous-échantillon par semaine pendant 10 ans. La population des unités de logement se trouve donc entièrement recensée au bout de dix ans, à l'exception des unités qui se sont ajoutées dans l'intervalle. Si le processus se poursuivait dans le temps, on pourrait obtenir à n'importe quel moment un portrait de la situation démographique des dix dernières années en agrégeant les données des 520 derniers sous-échantillons. C'est là une proposition intéressante et pleine d'imagination. Seulement, elle comporte des lacunes.

Kish propose un recensement par étapes où il n'y a pas de chevauchement d'échantillons d'une période à l'autre, sauf à la fin des dix années, lorsque le processus recommence. Avec un tel recensement, on obtiendrait à chaque semaine un grand échantillon transversal d'envergure nationale de même que des valeurs moyennes ou des valeurs agrégées pour chaque mois, chaque année et pour d'autres périodes. Toutefois, sans chevauchement d'échantillons, il sera relativement difficile de mesurer les variations qui surviennent d'une semaine à l'autre, d'un mois à l'autre ou même d'une année à l'autre dans les petites régions. On pourrait toujours prévoir des échantillons chevauchants, comme l'affirme Kish, mais cela accroitrait sensiblement le coût du recensement. La formule de recensement proposée par Kish permet néanmoins de mesurer des variations mais en l'absence d'échantillons partiellement chevauchants, les estimations de variations pour de petites régions seront affectées d'une forte erreur d'échantillonnage. Un des principaux objectifs des recensements décennaux est de produire des données pour les petites régions. Je crois qu'il est tout aussi important d'estimer avec précision les variations pour de petites régions que de produire des estimations agrégées pour des périodes précises. Bien que Kish le reconnait, il ne semble pas en faire cas.

Le sous-dénombrement deviendrait un problème particulièrement aigu dans un recensement par étapes. Comme la nécessité du recensement est généralement admise au sein de la population et que de grandes campagnes de publicité peuvent être faites pour les recensements, le taux de couverture pour les recensements a toujours été beaucoup plus élevé que celui pour les enquêtes par sondage, même dans les meilleures enquêtes (malgré cela, le sous-dénombrement demeure un problème pour les recensements). Le phénomène du sous-dénombrement répandu dans les enquêtes par sondage, même dans la Current Population Survey des E.-U., qui sert souvent de modèle. Il est donc difficile d'imaginer que l'on puisse soutenir l'intérêt du public pour un recensement par étapes au moyen d'une campagne de publicité permanente. Une autre difficulté que pose le recensement par étapes, à mon avis, est son coût vraisemblablement élevé. Kish le reconnaît mais ne semble pas en faire cas par la suite. Bien que je n'aie pas étudié d'estimations de coûts, je ne serais pas étonné de constater qu'un recensement par étapes, sur une période de dix ans, est beaucoup plus coûteux que des recensements quinquennaux classiques combinés à des enquêtes mensuelles relativement importantes qui visent à produire des estimations de variation ainsi que des données sur divers sujets pour les Etats et de grandes régions à l'intérieur des Etats. De plus, les recensements quinquennaux devraient être plus faciles à interpréter que les recensements par étapes et aussi plus utiles, en produisant des estimations pour petites régions pour des périodes précises ou de courts intervalles de temps au lieu de valeurs moyennes établies pour des périodes allant jusqu'à dix ans.

Attiré par la proposition de Kish concernant le recensement par étapes d'une part et résolu à étaler la charge de travail d'autre part, le Census Bureau est arrivé avec diverses propositions tournant autour de l'idée d'un recensement décennal réduit, accompagné de recensements en rotation. Le Census Bureau envisage diverses solutions à partir de l'idée de base, qui consiste à répartir le recensement des cinquante Etats sur dix ans suivant un ordre de rotation. C'est là une proposition originale qui vise à étaler la charge de travail sans engendrer les coûts exorbitants d'un recensement par étapes comme celui de Kish.

en collaboration avec Deming. Je faisais équipe avec Calvert Dedrick et Deming était associé à Philip Hausser et nous étions en relation constante avec Fred Stephan, qui agissait comme conseiller. Ensemble, nous avons élaboré cette importante étape dans l'application des techniques d'échantillonnage.

L'article de Barbara Bailar nécessite peu de commentaires additionnels. Nous avons en effet, moi, Bill Hurwitz et nos collègues, pris une part active aux projets de recherche qu'elle décrit si bien. J'aurais toutefois une petite correction à apporter à ses propos. Fienberg et Tanur considèrent à juste titre l'article qu'ont rédigé Hansen, Hurwitz, Marks et Mauldin en 1951 sur les modèles d'erreur de réponse comme la première publication à n'avoir jamais été produite sur le sujet, tandis que Bailar accorde ce crédit à un article de Hansen, Hurwitz et Bershad paru ultérieurement (1960). L'article de 1960 approfondissait les résultats de l'étude de 1951 et contenait des données empiriques obtenues par suite de l'application du modèle dans de vastes expériences de randomisation comme l'affectation aléatoire des recenseurs pour le recensement de 1950. L'analyse contenue dans cet article montrait jusqu'à quel point la corrélation d'erreurs due au travail des interviewers pouvait influencer les données de recensement pour petites régions. C'est principalement grâce à des mémoires parus antérieurement (où l'on présentait les mêmes résultats) et à des études connexes que l'on a décidé de recourir à la formule de l'auto-dénombrement pour les questionnaires abrégés du recensement de 1960. De plus, le questionnaire détaillé pouvait désormais être administré à un échantillon étendu de la population et non plus à la population entière, ce qui signifiait des économies appréciables, des délais de publication plus raisonnables et des données généralement plus précises. L'article de Bailar fait un excellent résumé de la question.

Par rapport à ce qui vient d'être dit, j'aimerais souligner l'apport remarquable de Bill (William N.) Hurwitz. Lui et moi avons formé une équipe dont le nombre de réalisations excède largement la somme de nos réalisations individuelles. De plus, je n'insisterai jamais trop sur les mérites de nos collègues, que nous avons recrutés et motivés et que nous avons dans une certaine mesure formés, et qui sont devenus les artisans de la recherche effectuée au Census Bureau en ce qui a trait à l'application de méthodes de sondage, de contrôle de la qualité et de recherche opérationnelle dans la conception et la réalisation de recensements et d'enquêtes par sondage dans des domaines variés. Parmi les collègues les plus célèbres, mentionnons les noms de Max Bershad, Joseph Daly, Leon Gifford, William Madow, Eli Marks, Harold Nisselson, Jack Ogus, Leon Pritzker, Joseph Steinberg, Benjamin Tepping, Joe Waksberg, Ralph Woodruff et d'autres. On attribue souvent à Morris Hansen les progrès qui ont été réalisés mais si ce n'avait été de Bill Hurwitz, en particulier, et de nos collègues, ces progrès n'auraient pas été possibles.

Je dois mentionner aussi que, de 1955 à 1968, année où j'ai laissé le Census Bureau, notre groupe a profité largement du concours et des conseils d'une équipe de statisticiens experts dirigée par Bill Cochran (William G. Cochran). Cette équipe comptait parmi ses membres Fred Stephan (Frederick F. Stephan) et Bill Madow (William G. Madow), qui en ont fait partie pendant toute la période en question, de même qu'Ivan Fellegi, de Statistique Canada, H.O. Hartley et d'autres, qui en ont fait partie pendant un certain temps. C'étaient toutes des personnes extrêmement compétentes. Toutefois, nous ne les considérons pas simplement comme des experts dont on cherche à obtenir l'avis et dont on applique les recommandations à la lettre. Notre relation avec eux était plutôt fondée sur l'échange. Nous discutons de problèmes précis et analysons toutes les étapes de l'élaboration d'un plan de sondage pour une enquête, une expérience ou un recensement en particulier. Nous avons bénéficié largement de leurs conseils et, réciproquement, ils ont tiré profit de leur relation avec nous.

L'article de Leslie Kish tranche avec les articles précédents en proposant de nouvelles méthodes de recensement pour l'avenir, notamment le recensement par étapes. Kish décrit par la même occasion diverses formes d'échantillons avec renouvellement complet.

Ces diverses formes d'échantillons (qu'ils aient ou non une portion chevauchante) ont actuellement plusieurs usages selon Kish et l'analyse qu'il en fait n'apporte rien de nouveau. Je

Cette expérience fit prendre conscience au groupe de la WPA de la nécessité d'une enquête permanente. Ils ont donc élaboré une enquête mensuelle sur la population active, qui présentait des aspects tout à fait inédits au point de vue du plan de sondage (mais aussi des lacunes qu'il a fallu corriger ultérieurement). Cette enquête n'était pas aussitôt rodée que les Etats-Unis se lancèrent dans la Seconde Guerre mondiale et les besoins de données changèrent du tout au tout. Il n'était plus question de chômage élevé mais de pénurie de main-d'œuvre. La WPA n'avait plus sa raison d'être et fut par conséquent supprimée et l'enquête dont elle avait la responsabilité fut confiée au Bureau of the Census; désormais, cette enquête devait servir essentiellement à mesurer l'effet de l'engagement militaire des Etats-Unis sur l'activité et l'emploi. Lorsque la responsabilité de l'enquête fut confiée au Bureau of the Census, nous avons remarqué des lacunes dans le plan de sondage original et avons élaboré des solutions en conséquence; c'est ainsi que nous avons introduit notamment l'échantillonnage avec PPT et d'autres innovations relatives au plan de sondage. Ces améliorations à l'enquête sur la population active (aujourd'hui la CPS, dont l'objet est beaucoup plus diversifié) ont eu un effet appréciable sur les méthodes d'échantillonnage et, chose plus importante encore, ont permis de satisfaire les besoins du pays en données de toutes sortes (données récentes sur la population active mais aussi données démographiques, sociales et économiques).

Fienberg et Tanur auraient pu aussi souligner les conséquences notables de la fusion des techniques de recensement et de sondage et de l'introduction de l'informatisation et de la lecture automatisée de repères dans les questionnaires de recensement. Depuis le recensement de 1960, les questionnaires qui servent à recueillir les données auprès des ménages sont assez sommaires. Des questionnaires plus complets sont distribués à des échantillons de la population recensee, qui sont tout de même en nombre respectable pour que l'on puisse obtenir des données utiles pour quelque 40,000 petites régions. Le recensement de 1960 a été aussi l'occasion d'introduire des méthodes d'autodénombrement. La décision d'introduire de telles méthodes a été motivée par l'application du modèle d'erreur de réponse auquel Fienberg et Tanur et par la réalisation d'études et d'expériences sur les erreurs de réponse et spécialement la correction entre erreurs de réponse attribuable au travail des recenseurs. Ces innovations étaient le résultat d'expériences majeures qui avaient été tentées avant et pendant le recensement de 1950 et dans des recensements ultérieurs, et d'autres expériences indépendantes. Une autre innovation importante a été FOSDIC (Film Optical Sensing Device for Input to Computers - appareil de lecture optique pour saisie de données), appareil conçu par le Bureau of Standards à la demande du U.S. Bureau of the Census pour lire les repères qui figurent dans les questionnaires du recensement; le Census Bureau comptait sur cet appareil pour éliminer tout le travail de perforation qui était requis durant les recensements. Ces innovations ont permis d'obtenir des données plus à point et généralement plus précises et de réduire le coût des recensements. Les perfectionnements sont devenus possibles dans ce domaine dès qu'on a pris conscience des problèmes que posaient les recensements "à grand déploiement" et qu'on a cherché à les résoudre en proposant par exemple l'utilisation de sondages et l'autodénombrement. Un autre facteur de progrès a été l'invention et la mise en service de l'ordinateur et du FOSDIC, deux champs d'action où le Census Bureau a fait figure de pionnier.

Vers la fin des années 1930, quelques membres de la haute direction du Census Bureau, appuyés de certains membres du Congrès, s'opposaient à ce que leur organisme fasse de l'échantillonnage. Il fallait respecter la tradition du recensement. L'utilisation de l'échantillonnage probabiliste dans le recensement de contrôle de 1937 qui accompagnait le recensement national des chômeurs a contribué largement à faire accepter l'échantillonnage au sein du U.S. Bureau of the Census; comme l'explique Bailar dans son article, l'échantillonnage devait être considéré comme une méthode tout à fait conforme aux objectifs du Census Bureau. Lors du recensement de la population de 1940, le Census Bureau a appliqué pour la première fois les techniques d'échantillonnage dans le but de recueillir des données supplémentaires (c'est-à-dire des données autres que celles normalement recueillies dans un recensement). A cette occasion, j'ai travaillé

J'aimerais ajouter un commentaire sur l'article de Rao et Bellhouse en ce qui a trait à l'estimation de la variance par la méthode "jackknife". Ils affirment que l'estimateur "jackknife" est non convergent pour des fonctions non lisses comme les quantiles, même dans le cas d'un échantillonnage aléatoire simple. Ils ont probablement raison d'affirmer cela si l'échantillonnage s'applique aux éléments qui servent d'unités d'analyse. Nous avons prouvé récemment de façon empirique qu'il était possible d'estimer avec précision la variance des médianes et (en l'occurrence) des 10^{ème} et 90^{ème} percentiles à l'aide de la méthode "jackknife" habituelle dans le cas d'un échantillonnage à plusieurs degrés où au moins deux unités du premier degré ou des combinaisons de celles-ci sont identifiées dans une strate (une étant éliminée et l'autre, reproduite de manière à obtenir un double). Nous supposons que la méthode "jackknife" donne des résultats efficaces dans ces circonstances parce que chaque grappe de dernier niveau associée à une unité du premier degré renferme un nombre appréciable d'unités ultimes d'échantillonnage. Nous estimons que cette méthode serait tout aussi efficace, bien que nous ne l'ayons pas démontrée, lorsque les échantillons redoublés sont formés au moyen d'une autre technique courante, selon laquelle on divise un échantillon aléatoire simple (ou aléatoire stratifié) en sous-échantillons aléatoires simples (ou aléatoires stratifiés – utilisant autant que possible les mêmes strates), qui sont par la suite retranchés un à un.

Fienberg et Tanur ont jeté un éclairage intéressant sur le rôle des institutions où ont pris forme les enquêtes par sondage. Je suis d'accord avec eux lorsqu'ils disent que la meilleure manière de comprendre plus à fond l'évolution des méthodes d'enquête est de connaître les institutions qui ont pour mandat de réaliser des enquêtes par sondage. La grande majorité des projets de recherche auxquels j'ai participé originaient des institutions et étaient souvent motivés par le besoin de résoudre les problèmes qui surgissaient au cours de l'exécution des programmes de ces institutions. Je veux, ici aussi, faire quelques commentaires sur certains aspects de la recherche à laquelle j'ai pris part.

Fienberg et Tanur affirment avec justesse que le plan de sondage de ce qu'on appelle maintenant la Current Population Survey ou CPS (et que l'on appelait auparavant l'enquête sur la population active) a joué un rôle déterminant dans l'évolution et l'application de la théorie des sondages et que ce rôle a de plus engendré d'autres progrès. Toutefois, ils laissent à entendre faussement que ce plan de sondage trouve son origine dans le recensement d'essai des chômeurs, réalisé entre 1933 et 1934 dans le cadre d'un programme de la Civil Works Administration (CWA) destiné à trois villes en particulier. Fienberg et Tanur semblent confondre le recensement d'essai réalisé par la CWA en 1933-1934 et le "recensement de contrôle" qui accompagnait le recensement des chômeurs de 1937. Comme ils l'indiquent eux-mêmes, c'est le recensement de 1937 qui a été élaboré conjointement par Dedrick, Hansen, Stouffer et Stephan et qui est à l'origine de la CPS. Il s'agissait d'un recensement à l'échelle nationale, dont la réalisation avait été confiée à l'Administration des postes. Le recensement de contrôle, qui était effectué par les facteurs, s'adressait à un échantillon probabiliste de routes postales prélevé à l'échelle nationale – chaque route de l'échantillon faisait l'objet d'un dénombrement complet. Ce recensement a été l'occasion d'appliquer de nouveaux principes d'évaluation de la population active et du niveau de chômage, fondés sur l'expérience des répondants au cours d'une semaine antérieure. C'était aussi la première fois que l'on utilisait un échantillonnage probabiliste aréolaire à la grandeur du territoire. Le recensement de contrôle avait pour but d'évaluer la précision des données du recensement des chômeurs de 1937 (comme l'explique Barbara Bailar dans l'article qui paraît dans ce numéro). Cette enquête par sondage a été très enrichissante et a permis de jeter les bases de l'enquête mensuelle sur la population active, qui allait devenir plus tard la Current Population Survey. Là encore, j'ai pris une part active à ces projets et Bailar le décrit bien. Stock, Frankel et Webb et d'autres à la Work Projects Administration (WPA) ont également participé à l'élaboration du recensement national et du recensement de contrôle. C'était à une époque où le chômage atteignait des proportions sans précédent et où il était devenu impératif d'en mesurer le niveau de façon continue.

Commentaires sur les articles de la section spéciale

MORRIS H. HANSEN¹

Ce sont là d'excellents articles que j'ai lus avec grand intérêt. Trois de ces articles portent plus particulièrement sur les aspects historiques et actuels de la question et abordent aussi brièvement l'avenir. L'article de Kish met l'accent sur le développement de nouvelles techniques et pose de fait les bases d'un tel développement. De par mon expérience personnelle et mon opinion sur la question, je vais tenter ici d'apporter quelques précisions sur le contenu des articles à caractère historique et de mettre en perspective la proposition de Kish voulant que les recensements par étapes remplacent désormais les recensements classiques.

Rao et Bellhouse font une analyse condensée mais utile de l'évolution des techniques de sondage. Après quelques remarques préliminaires, ils débudent leur rétrospective vers l'époque où j'ai commencé à m'intéresser aux recensements et aux enquêtes par sondage et à travailler à leur perfectionnement.

En tant que condensé, l'article de Rao et Bellhouse est ce qui peut se faire de mieux, les auteurs ne s'arrêtant pas outre mesure aux détails. Néanmoins, j'aimerais exprimer une opinion légèrement différente de la leur en ce qui a trait à l'échantillonnage avec probabilité proportionnelle à la taille ou à des mesures de taille (PPT). Ils ont raison de dire que nous (Hansen et Hurwitz) sommes à l'origine de la théorie de l'échantillonnage PPT avec remise s'ils entendent par cela que nous avons estimé des variances en supposant une forme approximative d'échantillonnage avec remise. Nous avons été incapables de résoudre la question de l'estimation de la variance dans le cas de l'échantillonnage avec PPT sans remise, question que ne devaient pas tarder à résoudre Horvitz-Thompson et d'autres. Cependant, sauf à quelques exceptions près, nous n'avons jamais proposé d'utiliser l'échantillonnage avec remise et n'y avons jamais eu recours non plus. Dans la pratique, nous avons appliqué l'échantillonnage avec PPT sans remise, soit en prélevant, à l'aide d'une méthode d'échantillonnage systématique, deux unités ou plus dans une strate où les unités étaient classées de façon aléatoire ou systématique, soit en prélevant une unité dans chaque strate. Les unités pour lesquelles la probabilité de sélection était élevée étaient choisies avec une probabilité égale à 1. Nous avons défini des estimateurs d'aggrégats et les fonctions correspondantes en utilisant l'inverse de la probabilité de sélection comme coefficient de pondération, à l'image de ce qui a été convenu d'appeler l'estimateur d'Horvitz-Thompson. Les estimateurs de la variance produisaient des estimations légèrement supérieures à la valeur réelle puisque, pour simplifier le problème, on avait supposé un échantillonnage avec remise. Cependant, des estimations de ce genre ne posent pas vraiment problème. On a souvent utilisé l'estimateur de la variance inter-grappe. Il s'agit là d'un estimateur approximatif de la variance d'une très grande simplicité et qui consiste en deux opérations: pondération (dans le cas d'un sous-échantillonnage) à l'intérieur des unités du premier degré et application des résultats à l'ensemble de l'unité, puis calcul de la variation entre ces unités (voir Hansen, Hurwitz et Madow, p. 257). Horvitz et Thompson ont bouleversé les fondements de l'estimation de la variance lorsqu'ils ont échantillonné plus d'une unité par strate avec des probabilités variées.

L'échantillonnage avec PPT présente les avantages que décrivent brièvement Rao et Bellhouse. De plus, il est d'une grande utilité dans les plans à plusieurs degrés, où les probabilités d'échantillonnage sont proportionnelles à des mesures de taille du premier au dernier degré. Les probabilités au dernier degré sont souvent établies de manière à obtenir des probabilités de sélection globales uniformes pour les unités ultimes.

¹ Morris H. Hansen, Westat, 1650 Research Boulevard, Rockville, MD, 20850, U.S.A.

- JENSEN, P. (1983). Towards a Register-Based Statistical System – Some Danish Experiences. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.
- JENSEN, P. (1987). The Quality of Administrative Data from a Statistical Point of View: Some Danish Experience and Consideration. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs et M.P. Singh (éds.) Statistique Canada, Ottawa.
- JOBE, J.B., et MINGAY, D.J. (1990). Cognition and Survey Measurement: History and Overview. *Applied Cognitive Psychology*, à paraître.
- KUHN, T.S. (1970). *The Structure of Scientific Revolutions*. Second Edition, Enlarged, The University of Chicago Press, Chicago.
- MEYER, B. (1990). The Tax System: Comparisons of Demographic, Labour Force and Income Results for Individuals and Families. Division des données régionales et administratives, Statistique Canada.
- PLATEK, R., RAO, J.N.K., SÄRNDALE, E.E., et SINGH, M.P. (1987). *Small Area Statistics*, New York: Wiley-Interscience.
- PODOLUK, J. (1987). Administrative Data as Alternative Sources to Census Data. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs et M.P. Singh (éds.), Statistique Canada, Décembre 1988, Ottawa, 273-290.
- REDFERN, P. (1987). A Study of the Future of the Census of Population: Alternative Approaches Eurostat Theme 3 Series C. Luxembourg: Office for Official Publications of the European Communities.
- REDFERN, P. (1989). Population Registers: Some Administrative and Statistical Pros and Cons. *The Journal of the Royal Statistical Society, series A (Statistics in Society)*, 152, 1-41.
- ROYCE, D., et DREW, J.D. (1988). Address Register Research: Current Status and Future Plans. 1991 Research and Testing Project, 1991 Census, Statistique Canada, Ottawa.
- SCHREUREN, F., ALVEY, W., et KILSS, B. (1990). Paradigm Shifts: Administrative Records and Census-Taking.
- STATISTIQUE CANADA (1990). Research papers and reports. Bibliography, Division des données régionales et administratives, Ottawa, Ontario. (inédit).
- TEGELS, R., et CAHOON, L.S. (1982). The Redesign of the Current Population Survey: The investigation into alternate rotation plans. *Proceedings of the American Statistical Association, Survey Research Methods Section*.
- U.S. BUREAU OF THE CENSUS (1989). *200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC.
- VERMA, R.B.P., et RABY, R. (1989). Utilisation des fichiers administratifs pour estimer la population au Canada. *Techniques d'enquête*, 15, 271-280.

- ALVEY, W. et SCHEUREN, F. (1982). Background for an Administrative Record Census. 1982 *American Statistical Association Proceedings, Social Statistics Section*, 137-146.
- ANDERSON, M. (1990). 'According to their respective numbers . . . ?' For the twenty-first time. *Chance*, 3, 12-18.
- BAILLAR, B. (1990). Contributions to Statistical Methodology from the Federal Government. *Survey Methodology*, 16.
- BARKER, J.A. (1988). *Discovering the Future: The Business of Paradigms*.
- BATES, N.A., et DEMAYO, T.A. (1989). Using Cognitive Research Methods to Improve the Design of the Decennial Census Form. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 267-285.
- BOUNPANE, P. (1988). A Sample Census: A valid alternative to a complete count census? 46th Session of the International Statistical Institute.
- BROWNE, D.L. (1989). U.S. Bureau of the Census: Facing the future labor shortage. *Asian and Pacific Population Forum*, 3, 4.
- BUTZ, W. (1985). Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities. *Journal of Business and Economic Statistics*, 393-395.
- CITRO, C., et COHEN, M.L. (eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. National Academy Press, Washington, DC.
- DIPPO, C. (1987). A Review of Statistical Research at the U.S. Bureau of Labor Statistics. *Journal of Official Statistics*, 3, 289-297.
- DREW, J. D. (1989). Address Register Development and its possible future role in Integration of Census, Survey and Administrative Data. Présenté au U.S. Bureau of the Census/Statistics Canada Interchange. (Inédit).
- FELLECI, J.P. (1981). Discussion de l'article par Leslie Kish intitulé "Population Counts from Cumulated Samples." *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau. An Analysis, Review and Response*, Congressional Research Service, the Library of Congress.
- FIENBERG, S. (1990). An Adjusted Census in 1990? An Interim Report. *Chance*, 3, 19-21.
- FIENBERG, S., et TANUR J. (1989). Combining Cognitive and Statistical Approaches to Survey Design. *Science*, 243, 1017-1022.
- GRACE, J.W. (1989). The Use of Administrative Records for Social Research. Atelier Statistique Canada, le 12 décembre 1989, Ottawa, Ontario.
- HAMMOND, R.B. (1990). The 1990 Decennial Census: An Overview. *Conference Proceedings, Advanced Computing for the Social Sciences*, Oak Ridge National Laboratory et U.S. Bureau of the Census, 10-12 avril, 1990, Williamsburg, Virginia.
- HERriot, R., BATEMAN, D.V., et McCARTHY, W. F. (1989). The Decade Census Program - New Approach for Meeting the Nation's Needs for Sub-National Data. *American Statistical Association Proceedings, Social Statistics Section*.
- HUGGINS, V., et FAY, R. (1988). Use of Administrative Data in SIPP Longitudinal Estimation. *American Statistical Association Proceedings, Section on Survey Research Methods*, 354-359.
- IRWIN, R. (1984). Feasibility of an Administrative Records Census in 1990. Special report on the use of administrative records, committee on the use of administrative records in the 1990 Census, rapport inédit du Census Bureau.
- JABINE, T.B., et SCHEUREN, F. (1985). Goals for Statistical Uses of Administrative Records: The Next Ten Years. *Journal of Business and Economic Statistics*, 380-391.
- JABINE, T.B., et SCHEUREN, F. (1987). Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going? *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs et M.P. Singh (eds.), Statistique Canada, Décembre 1988, Ottawa, 43-72.

- Aux États-Unis, le Census Bureau a commencé à étudier des méthodes qui pourraient être employées pour remplacer les recensements conventionnels (Bounpane 1988). Malheureusement, les recherches nécessaires pour étudier la possibilité d'utiliser des fichiers administratifs à cette fin viennent à peine de commencer. Il reste à voir si le Census Bureau trouvera une meilleure approche que l'emploi des fichiers administratifs et des échantillons avec renouvellement complet (Browne 1990). Cependant, quelles que soient les autres solutions de rechange qu'ils étudient, ils devront certainement examiner l'utilisation des fichiers administratifs comme remplacement partiel pour les données intégrales conventionnelles. On trouve, dans Scheuren, Alvey et Kilss 1990, un calendrier de recherche préliminaire qui met à jour des idées déjà présentées.

Kish a raison de dire que, compte tenu des propositions radicales dont lui et moi discutons, nous ne pouvons donner de réponse catégorique. Comme lui, je crois que, dans la majorité des cas, un changement par rapport aux méthodes utilisées pour réaliser les recensements conventionnels permettrait d'améliorer "la valeur des composantes de la variance". "Cependant, des études théoriques aussi bien qu'empiriques seront nécessaires pour résoudre cette question." Bien entendu, pour un changement aussi considérable que celui qui est proposé ici, l'"équilibre" qui doit être atteint dépasse de beaucoup le fait de ne tenir compte que des composantes de la variance (et du biais). Kish reconnaît cette situation de nombreuses façons dans son article. Cependant, on devrait insister davantage sur le fait que certains aspects, au moins, des changements de paradigme que l'on examine pourraient aller au cœur du contrat social qui existe entre les organismes statistiques nationaux et les personnes que ces organismes ont mission de servir. Par exemple, on trouve dans la Constitution des E.-U. l'obligation de "dénombrer" la population tous les dix ans. L'utilisation de fichiers administratifs ou de recensements par étapes s'accorderait-elle avec ce "paradigme constitutionnel"? On pourrait peut-être commenter par adoption une définition plus étendue de "dénombrer".

Un autre exemple où des questions relatives au contrat social se présentent est celui de la mesure dans laquelle on pourrait considérer que l'utilisation accrue de données administratives existantes (ou de données additionnelles) à des fins statistiques pourrait être vue comme un accroissement fâcheux de l'intrusion de l'État dans la vie privée de ses citoyens (Grace 1989). Si légitimes que soient ces inquiétudes relatives à l'"intrusion", il ne semble pas y avoir de preuves, dans un contexte nord-américain, du moins, qu'elles constituent un obstacle insurmontable. Au contraire, il n'y a presque pas eu de réactions défavorables du public quand, aux E.-U., on a ajouté des éléments aux fichiers administratifs à des fins statistiques (p. ex., des renseignements sur l'adresse domiciliaire avec les déclarations d'impôt de 1972, 1974 et 1980). À ma connaissance, cette question n'a pas encore été soulevée directement au Canada, du moins au niveau fédéral.

En résumé, pour apporter des changements des genres dont Kish discute, il faut, comme il le signale, réaliser beaucoup de recherches scientifiques. L'étude des techniques qui devront être employées pour la mise en application de ces changements constituera un travail plus considérable encore. Finalement, ces questions dépassent notre profession et il est fort possible qu'elles se règlent dans d'autres arènes. Où que les décisions soient prises, il nous incombe, à titre de statisticiens, de présenter le débat sous forme d'options réalisables. Kish nous a fait parcourir beaucoup de chemin dans cette direction et il mérite toutes nos félicitations.

BIBLIOGRAPHIE

- ALVEY, W. et KILSS, B. (éds.) (1990). *Statistics of Income and Related Administrative Record Research*. U.S. Department of the Treasury, Internal Revenue Service. Voir aussi Kilss, Beth et Alvey, Wendy (éds.) (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, vols. 1 et 2, U.S. Department of the Treasury, Internal Revenue Service.

comprendre toutes les différences d'ordre conceptuel entre la signification des termes quand ils sont employés dans des enquêtes ou tirés de fichiers administratifs. De toute façon, il se peut que nous ne disposions pas de ce que nous pensons avoir (Bates et DeMaio 1989). Dans tous les cas, il existe déjà un ensemble considérable de recherches cognitives auquel on peut faire appel (voir p. ex., Dippo 1987; Fienberg et Tanur 1989; Jobe et Mingay 1990).

Kish est près de la réalité quand il dit que les fichiers administratifs "ne pourront répondre aux exigences de la société moderne en ce qui concerne les sources de données." Bien entendu, ces exigences semblent insatiables. Même si elles ne l'étaient pas, les fichiers administratifs ne seront jamais aussi flexibles et sensibles que les enquêtes. Cependant, quand des fichiers (y compris des fichiers partiels comme ceux qui existent aux E.-U.) sont reliés à des données d'enquête, ils peuvent être extrêmement importants comme variables auxiliaires pour trouver des estimations directes et améliorées relatives à une enquête nationale – et même à une enquête infranationale. La recherche effectuée par le U.S. Census Bureau dans le cadre du Survey of Income and Program Participation sur l'utilisation des données du Internal Revenue Service afin d'améliorer la précision des estimations nationales est un bon exemple récent (Huggins et Fay 1988). Il faudrait encore trouver des estimations régionales indirectes (p. ex., des estimations synthétiques) pour les variables qui ne figurent pas dans les fichiers administratifs (Platek, Rao, Särndal et Singh 1987). Cependant, les fichiers pourraient fournir une source d'indicateurs symptomatiques précieux.

Conclusions

L'exposé que fait Kish afin que l'on étudie un "changement de paradigme" pour la réalisation des enquêtes semble irrésistible, du moins dans le pays nantis comme le Canada et les E.-U. La solution de rechange qu'il propose, soit les recensements par étapes, est probablement trop dispendieuse pour qu'on la mette en application afin de remplacer complètement un recensement. Cependant, les échantillons avec renouvellement complet sont très prometteurs, s'ils peuvent être intégrés aux opérations des enquêtes permanentes actuelles des programmes statistiques nationaux du Canada et des E.-U. De tels échantillons pourraient fournir un lien nécessaire afin de répondre aux besoins d'estimations pour petites régions auxquels il serait impossible de répondre autrement. Leur emploi comme substitut (partiel) pour les échantillons de personnes qui doivent remplir le questionnaire complet du recensement est moins prometteur mais quand même possible.

Il se peut que Kish soit trop pessimiste à propos des fichiers administratifs. Cependant, la situation au Canada diffère de celle que l'on retrouve aux E.-U.:

- Au Canada, on peut déjà penser à combiner des échantillons avec renouvellement complet et des fichiers administratifs comme méthode à employer pour remplacer les recensements conventionnels. Cela ne veut pas dire qu'il ne reste pas d'énormes défis pratiques à relever. Cependant, la partie du recensement du Canada pour laquelle on recueille des données intégrales pourrait être réalisée en utilisant les fichiers administratifs comme point de départ, on pourrait leur ajouter une enquête à grande échelle afin de mesurer le sous-dénombrement et, potentiellement, pour corriger cette situation. L'échantillon de 20% des ménages qui, au Canada, doivent répondre au questionnaire complet du recensement pourrait, du moins en partie, être remplacé par un échantillon avec renouvellement complet. Le contenu du questionnaire complet utilisé pour le recensement est beaucoup plus riche que dans le cas des enquêtes-ménages, mais les différences au niveau du contenu pourraient être compensées au moyen de questions additionnelles ajoutées, à intervalles réguliers, aux enquêtes permanentes. Les échantillons avec renouvellement complet pourraient aussi permettre d'aborder directement les problèmes de couverture relatifs à l'emploi des fichiers administratifs, particulièrement à des fins de calibration pour tenir compte des changements dans ces fichiers entre les recensements.

dans ce que certains, au moins, appelleraient une position intellectuelle de plus en plus stérile (Fienberg 1990). Le point de vue adopté par les dirigeants du Bureau of the Census fait qu'il leur est très difficile de voir toute autre possibilité, comme une approche qui utiliserait (en partie) des fichiers administratifs et pour laquelle il faut admettre au départ que des rajustements devront être apportés.

La situation est différente au Canada. Depuis le début des années 80, Statistique Canada a assemblé un bon nombre des modules nécessaires pour réaliser un recensement à partir des fichiers administratifs (voir p. ex., Drew 1989; Podoluk 1987; Verma et Raby 1989). Bien qu'il reste beaucoup de travail à effectuer, un tel changement pourrait même se produire aussi tôt qu'en 1996. La couverture du système fiscal canadien est très élevée et elle croît. En 1987, par exemple, on a estimé que la couverture était d'environ 94% – c.-à-d., environ 3% de moins que la couverture de 96,8% atteinte lors du recensement du Canada de 1986. En 1991, la couverture assurée par les fichiers du système fiscal devrait atteindre environ 97% ou plus. En outre, il est probable que la couverture assurée par les fichiers administratifs croîtra encore au cours des années 90.

Kish fait part de ses inquiétudes à propos du fait que les fichiers administratifs, même une fois que leur qualité et que la couverture qu'ils assurent seront devenues adéquates, "ne contiendront jamais d'autre chose qu'un minimum de variables démographiques: chiffres de population, âge, sexe et quelques autres éléments." Une observation immédiate que l'on peut faire à propos sa remarque est le fait que les recensements conventionnels permettent de recueillir eux-mêmes, bien *par* d'autres renseignements, du moins pour les rubriques servant à recueillir des données intégrales. Il est aussi évident que, bien que les variables contenues dans les fichiers administratifs diffèrent de celles qui sont recueillies dans le cadre d'un recensement traditionnel, il y en a beaucoup plus qui sont *déjà* disponibles que Kish ne le réalise (voir p. ex., Meyer 1990; Alvey et Scheuren 1982).

Le besoin d'insister sur le fait que la proposition visant à utiliser les fichiers administratifs pour la réalisation des recensements *ne* prévoit *pas* que les fichiers administratifs doivent être utilisés tels qu'ils sont actuellement est même plus important que toute comparaison du contenu actuel des rubriques. *Les fichiers administratifs devront être modifiés*. Selon moi, un optimisme limité est justifié pour ce qui est de réaliser les modifications nécessaires. Cependant, il ne fait aucun doute, que l'on ne peut s'attendre à ce que les fichiers administratifs pourront renfermer des données relatives exactement aux mêmes concepts qui sont mesurés maintenant dans les recensements et dans les enquêtes. De plus, il est presque certain que des efforts spéciaux devront être faits, à l'aide des techniques utilisées actuellement pour réaliser les recensements, afin de dénombrer séparément certains groupes. Les efforts déployés dans le cadre du recensement de 1990 aux E.-U. pour dénombrer les sans abri constituent un exemple de ce genre.

Les recensements ainsi que les fichiers administratifs ont tous des limitations inhérentes. Les différences inévitables au niveau conceptuel seront des obstacles importants lors de tout passage d'une de ces sources de données à l'autre. La faisabilité du point de vue administratif constitue une autre question; cependant, il se peut que certains concepts difficiles à reproduire du recensement (p. ex., les ménages) ne soient pas aussi importants pour le processus de mesure que ce n'était le cas auparavant.

Pour certaines utilisations, les changements d'ordre méthodologique (p. ex., le passage du recensement aux fichiers administratifs) pourraient bien être accompagnés d'une modification parallèle dans les concepts sous-jacents mesurés. La signification de certains concepts, ainsi que notre capacité à les mesurer (p. ex., les familles) peut être modifiée ou étendue. Nous devons aussi établir dans quelle mesure les répondants répondent aux questions des enquêtes de la même façon qu'ils remplissent les formulaires administratifs qui peuvent avoir un effet direct et réel sur leur vie.

Au cours des dernières années, l'utilisation de nouveaux outils provenant du domaine de la psychologie cognitive a permis d'améliorer les méthodes traditionnelles utilisées pour les enquêtes. Ces outils employés dans les recherches cognitives pourraient être utilisés pour

Tant au Canada qu'aux E.-U., il se peut que l'on doive régler le problème des coûts unitaires vraisemblablement plus élevés pour un échantillon avec renouvellement complet en apportant des modifications aux procédures employées pour les enquêtes: comment les segments sont-ils inscrits sur les listes? (Royce et Drew 1988); comment la prise de contact avec les ménages est-elle réalisée?, etc. Où est-il écrit, par exemple, qu'il faut effectuer une interview sur place avant d'avoir recours à d'autres méthodes de collecte des données?

Le défi à relever afin de conserver des tailles effectives des échantillons égales pour le niveau principal et de changer les données obtenues actuellement à l'aide des enquêtes permanentes (voir p. ex., Tegels et Cahoon 1982) est considérable. Il se peut, de plus, que l'on doive faire certains compromis relativement au contenu de base des questionnaires complets du recensement aux quels les échantillons du recensement doivent répondre actuellement. Malgré ces défis, on peut-être à cause d'eux, les idées de Kish sur les échantillons avec renouvellement complet méritent qu'on leur porte une attention sérieuse et continue et elles devraient être soumises à des essais pratiques approfondis.

Fichiers administratifs

Avec l'épanouissement des méthodes scientifiques d'enquêtes par sondage au cours des années 40 (Bailar, 1990), l'utilisation des fichiers administratifs à des fins statistiques est devenue relativement moins importante dans de nombreux programmes statistiques nationaux. Cependant, au début des années 80, du moins dans les pays nantis, le mouvement du pendule avait commencé son retour. Kish reconnaît cette tendance et il cite, à juste titre, Philip Redfern, qui a été le principal chroniqueur de ce phénomène à l'échelle internationale (Redfern 1987). Bien que les Danois semblent avoir été le plus loin dans ce domaine (Jensen 1983 et 1987), des efforts importants ont été faits au Canada (voir p. ex., Statistique Canada 1990) et certains ont même été réalisés aux E.-U. (voir p. ex., Alvey et Kilss 1990).

On trouve un bon résumé de la plupart des obstacles clés à l'utilisation accrue des fichiers administratifs pour la réalisation des recensements dans Redfern (1989), ainsi que dans l'analyse approfondie publiée avec cet article. Les obstacles en matière de perception qui touchent les citoyens (p. ex., en Allemagne) sont mentionnés comme des problèmes. Les obstacles d'ordre psychologique relatifs au service statistique national peuvent, cependant, avoir une importance égale ou même supérieure. Les principaux "changements de paradigmes" scientifiques ont généralement ce problème (Kuhn 1970). Cela a certainement semblé faire partie de la raison fournie pour la réception donnée à la proposition (que j'ai faite en 1980) afin d'étudier la faisabilité de faire des fichiers administratifs une partie intégrante du recensement de la population (Census of Population) des E.-U. Bien qu'un aperçu de cette proposition ait finalement été présenté lors des réunions de 1982 de l'American Statistical Association (Alvey et Scheuren 1982), il semble, sauf pour certaines exceptions assez limitées (voir p. ex., Irwin 1984; Citro et Cohen 1985), que le Census Bureau ait manifesté remarquablement peu d'intérêt sérieux pour cette proposition.

Je me contenterai de dire qu'aux E.-U., on a entrepris très peu des recherches nécessaires. Cela est vrai, en dépit d'efforts ininterrompus visant à donner de l'importance à la proposition (Jabine et Scheuren 1985 et 1987) et à ce qu'on en discute beaucoup (Butz 1985). Je dois donc, malheureusement, admettre que Kish a probablement raison de dire qu'aux Etats-Unis, du moins pour l'an 2,000, "... il serait bien étonnant [que les fichiers administratifs] remplacent les recensements. . . ."

Le recensement décennal de 1990 aux E.-U. aurait pu être utilisé pour faire certaines recherches portant sur l'emploi de fichiers administratifs pour remplacer certaines parties du recensement et ainsi prouver ou réfuter l'utilité de cette méthode. On ne peut que faire des spéculations sur la raison pour laquelle cela ne s'est pas produit. Il est fort possible qu'un cas de "paralyse de paradigme" ait été un facteur qui a contribué à cette situation (Barker 1988). Il semble que la controverse visant à déterminer si l'on doit ou non rajuster les "chiffres" du recensement, qui dure depuis littéralement des décennies, semble avoir figé le U.S. Bureau of the Census

forme la plus pure, l'espace et le temps deviennent une seule dimension et le contenu demeure fixe, de sorte qu'à la fin de la décennie, nous avons obtenu des renseignements cumulatifs sur tout le pays pour un ensemble donné de rubriques.

L'avantage principal d'un recensement par étapes est qu'il permet d'éviter le problème de l'obsolescence des données tant au niveau national qu'aux principaux niveaux infranationaux. Pour les petites régions géographiques, cependant, il n'y aurait encore, bien entendu, qu'une observation par décennie. Contrairement à ce qui se produit dans le cas d'un recensement conventionnel, les comparaisons entre les petites régions géographiques seraient très difficiles à interpréter parce que les données sont recueillies à des moments différents (Fellegi 1981).

Pour un recensement ou une enquête par étapes, les coûts unitaires pourraient être plus élevés, comme Kish le fait remarquer, que pour un dénombrement plus conventionnel (en effet, toutes choses étant égales par ailleurs, il se peut même que ces coûts soient plus élevés que ceux des enquêtes réalisées actuellement). À une époque où les ressources sont fixes ou diminuent, il pourrait donc ne pas être possible d'effectuer un "dénombrement" complet au cours de chaque décennie, même si le contenu du recensement était réduit considérablement. Il semblerait que l'attrait principal des recensements par étapes ne soit pas tant le fait qu'on puisse les utiliser à la place des recensements conventionnels, mais, disons, dans le cadre d'une stratégie visant à relier la réalisation des recensements avec les enquêtes permanentes et les estimations relatives à la population de régions locales pour les années intercensitaires (Herriot, Bateman et McCarthy 1989).

Les E.-U. tout comme le Canada, ont recours à des enquêtes mensuelles pour estimer les caractéristiques nationales de la population active (et certaines caractéristiques infranationales). Au Canada, l'Enquête sur la population active (EPA) est menée auprès de 64,500 ménages, c'est-à-dire qu'elle porte sur 0,67% de l'ensemble de la population canadienne chaque mois. (traduction) "Compte tenu du schéma de renouvellement utilisé pour l'EPA, l'échantillon de 0,67% par mois se ramène à un échantillon de 6,7% de ménages uniques au cours d'une période de 5 ans" (Drew 1989). Dans le contexte canadien, du moins, il se peut que la conjecture de Kish soit juste. On pourrait concevoir un véhicule utilisé pour les enquêtes-échantillons avec une certaine réduction dans le chevauchement entre les ménages d'un mois à l'autre, ce qui permettrait d'obtenir un bon nombre des avantages que Kish a exposés pour un échantillon avec renouvellement complet, tout en répondant aux besoins de renseignements que les enquêtes-ménages permanentes satisfont actuellement (Drew 1989). Cet échantillon ne remplacerait pas les données intégrales du recensement lui-même, mais il pourrait remplacer partiellement l'échantillon des personnes qui, au Canada, répondent au questionnaire complet (20% des ménages).

Parce que la population des États-Unis est environ 10 fois plus grande que celle du Canada, les compromis entre les échantillons avec renouvellement complet et la couverture globale du pays ne sont pas aussi intéressants qu'ils le sont au Canada. Par exemple, la Current Population Survey (CPS) des E.-U. menée auprès de 60,000 ménages ne porte que sur 0,06% de la population totale des E.-U. chaque mois. Même si les données qu'elle permet de recueillir étaient accumulées sur une décennie complète (mais sans changement dans le schéma de renouvellement utilisé), la CPS ne porterait que sur environ 1% de tous les ménages aux E.-U. Cette taille est loin d'être comparable à l'échantillon global de 16% des ménages qui doivent remplir un questionnaire complet dans la cadre du recensement des E.-U. de 1990.

Pour rapprocher la population de l'échantillon avec renouvellement complet de celle de l'échantillon utilisé pour le recensement décennal des E.-U. de 1990, il faudrait apporter des modifications importantes, comme celles que Kish nous demande de considérer, au schéma de renouvellement utilisé pour la CPS. Il se pourrait aussi que l'on doive remanier d'autres enquêtes du U.S. Census Bureau, si l'on visait à réaliser une substitution, même partielle. De plus, en dépit de ces modifications, l'échantillon résultant pour toute la décennie ne représenterait quand même qu'un faible pourcentage de la population totale des E.-U. – peut-être, au mieux, entre 2 et 3%, si l'on suppose que les besoins en ressources et autres restaient essentiel-

lement fixes.

la différence dans les coûts. Tant le recensement de 1960 que celui de 1990, par exemple, ne posaient, à chaque recensement, que 7 questions sur la population (U.S. Bureau of the Census 1989). Le questionnaire complet utilisé pour le recensement de 1960 contenait 35 questions et il devait être rempli par 25% de la population. Pour 1990, 16% des ménages américains devaient répondre au questionnaire complet qui comprenait 33 questions.

- La situation est semblable au Canada, pour ce qui est des coûts de la réalisation du recensement. Par exemple, le budget du recensement du Canada de 1991 prévoit environ \$9.50 CAN par personne. Comme pour le recensement des E.-U., tous les recensés doivent répondre à seulement 7 rubriques, cependant légèrement différentes, portant sur la population. Comme pour le recensement de 1990 aux E.-U., tout le monde doit répondre à des questions sur le logement (2 au Canada et 7 aux E.-U.). Au Canada, un questionnaire complet sera utilisé pour 20% des ménages en 1991. Ce questionnaire comprend 45 rubriques. Le recensement du Canada de 1961 était fort différent de celui qui est prévu pour 1991, il est donc difficile de faire des comparaisons de coûts significatives. Toutefois, si l'on revient 30 ans en arrière au Canada, il est évident que les coûts semblent avoir suivi la même tendance à long terme que celle observée aux E.-U.; cependant, en ce qui concerne les deux ou trois derniers recensements, les coûts per capita se sont stabilisés – ils ont même légèrement régressé.

Le U.S. Census Bureau a étudié le coût croissant des recensements conventionnels et il a conclu qu'il se peut qu'un changement important soit nécessaire (Browne 1990). Les coûts de la main d'œuvre ont augmenté de façon appréciable au cours des dernières décennies, tant au Canada qu'aux E.-U. Les améliorations technologiques n'ont pas été suffisantes pour contrebalancer ces coûts, bien que certaines comme TIGER et l'ITAO, soient prometteuses. L'attention plus grande accordée, aux E.-U., à l'amélioration de la couverture de la population est un autre facteur important (Anderson 1990). La collaboration que le public apporte au recensement a aussi diminué, du moins comme le témoigne le taux de réponse postale plus faible que prévu pour le recensement de 1990. (Au Canada, les fluctuations de la collaboration du public ne montre aucune tendance nette.)

Les coûts croissants ne constituent pas le seul problème important qui se présente en rapport avec la réalisation des recensements conventionnels. Le taux croissant d'obsolescence des données recueillies a peut-être, comme Kish le fait remarquer, une importance encore plus grande. La combinaison des coûts qui augmentent et de l'obsolescence croissante des données a eu pour effet de réduire continuellement et de façon spectaculaire le ratio coûts-avantages pour les recensements conventionnels.

Afin d'obtenir des données régionales plus fréquentes, certains pays ont commencé à réaliser des recensements quinquennaux. Au Canada, par exemple, c'est en 1956 qu'un tel recensement a été réalisé pour la première fois à l'échelle nationale. Cependant, des restrictions budgétaires ont presque entraîné l'annulation du recensement de 1986 au Canada, et l'on est réellement inquiet à propos de la façon dont le recensement y sera réalisé en 1996, on se demande même s'il y en aura un. Bien qu'une loi autorisant la réalisation d'un recensement quinquennal ait aussi été adoptée aux E.-U., on n'a jamais disposé des fonds nécessaires.

Recensements par étapes

Comme Kish le fait remarquer à juste titre, pour réaliser des recensements conventionnels, on doit nécessairement sacrifier l'actualité des données ainsi que le contenu des rubriques (au niveau des données intégrales) afin d'obtenir les détails géographiques complets et une couverture élevée au niveau de la population.

Le 'recensement par étapes' est une des possibilités que Kish nous demande de considérer. Sa proposition prévoit le dénombrement d'un pays par échantillonnage, au cours d'une période de dix ans, de façon à ce qu'à la fin toutes les régions aient été dénombrées. Dans sa

COMMENTAIRES FRITZ SCHEUREN¹

Les ouvrages statistiques ont négligé l'idée des échantillons cumulatifs. Dans plusieurs articles déjà publiés ainsi que dans le présent article, Leslie Kish a tenté de corriger la situation. Tous jours ouvert et pratique, il fait un exposé convaincant et irrésistible en faveur de la réalisation de travaux additionnels portant sur les questions relatives à la conception et à l'analyse soulevées par l'aggrégation.

Ses écrits sont tellement terre à terre que les lecteurs peuvent manquer le fait que Kish ne préconise pas seulement d'ajouter quelques éléments mineurs à la quantité déjà considérable de plans de sondage et de méthodes d'estimation dont nous disposons. Il nous demande d'examiner de très près la topologie des compromis, en matière d'espace, de temps et de contenu, relatifs aux enquêtes – plus particulièrement aux recensements. En fait, Kish semble préconiser ce que l'on pourrait appeler un "changement de paradigme" dans la réalisation des recensements, du moins dans les pays nantis comme le Canada et les E.-U.

Le mot "paradigme" mérite une certaine élaboration (Barker 1988). Un paradigme est une façon de penser, puis d'agir, une combinaison d'opinions et de comportements, une façon de voir la réalité et d'utiliser cette perception pour accomplir quelque chose. Les paradigmes sont des choses courantes – le chemin que nous prenons pour nous rendre au travail en constitue un exemple élémentaire. Selon cette définition, on pourrait dire de la réalisation des recensements conventionnels qu'il s'agit d'un paradigme scientifique et technique important.

Tant que nos paradigmes ne nous causent pas de problèmes, nous avons tendance à ne pas les changer. Cependant, il arrive que les paradigmes s'effondrent et qu'on doive les remplacer. Le pont s'écroule et nous devons trouver un autre itinéraire pour nous rendre au travail. Comme Kuhn l'a fait remarquer dans son ouvrage précurseur sur la structure des révolutions scientifiques (Kuhn 1970), L'exemple le plus fameux d'une telle situation est peut-être la révolution, dans les opinions des astronomes, qui s'est produite quand le système géocentrique de l'univers de Ptolémée a été remplacé par le système copernicien dans lequel la terre tournait, avec les autres planètes, autour du soleil.

Kish, dans son article, soutient qu'il existe des problèmes importants en rapport avec le paradigme qui s'applique à la réalisation des recensements traditionnels. Il examine ensuite dans le détail deux méthodes de remplacement: les recensements par étapes et les fichiers administratifs. Mon objectif ici sera de compléter et occasionnellement d'équilibrer la présentation de Kish sur ces sujets.

Réalisation des recensements conventionnels

Les recensements conventionnels, comme ceux du Canada et des E.-U., continuent de faire beaucoup de choses très bien. En fait, nous ne disposons actuellement de rien qui pourrait les remplacer adéquatement; néanmoins, l'opinion de Kish qui voudrait qu'on leur apporte au moins certains changements semble irrésistible. Les coûts croissants constituent un facteur important. De nombreuses améliorations ont été apportées à la réalisation des recensements au cours de ce siècle; néanmoins, tant au Canada qu'aux E.-U., les coûts totaux et même les coûts par personne ont augmenté considérablement:

- Dans le budget du recensement décennal de 1990 aux E.-U., on prévoit environ \$10 US par personne. Même si l'on tient compte de l'inflation, cela signifie que les dépenses par habitant ont quadruplé depuis 1960. Les différences entre les deux recensements au niveau des rubriques sont petites et, essentiellement, elles ne constituent pas un facteur qui explique

¹ Fritz Scheuren, Director, Statistics of Income Division, Internal Revenue Service. Les opinions exprimées dans cet article sont celles de l'auteur et n'engagent nullement le Internal Revenue Service.

- KISH L. (1981). *Using Cumulated Rolling Samples*. Washington: Congressional Research Office, 80-528-0.
- KISH L. (1987). *Statistical Designs for Research*. New York: John Wiley and Sons.
- KISH L. (1988). Plans de sondage à usages multiples. *Techniques d'enquête*, 14, 19-33.
- KISH L. (1989). Developing statistics in China. *Journal of Official Statistics*, 5, 157-69.
- KISH L., LOVEJOY, W., et RACKOW, P. (1961). A multi-state probability sample for traffic surveys. *Proceedings of the Section on Social Statistics, American Statistical Association*, 227-230.
- MOONEY, H.W. (1956). *Methodology in Two California Health Surveys*, San Jose (1952) and Statewide (1954-55). U.S. Public Health Monograph No. 70.
- NATIONAL CENTER FOR HEALTH STATISTICS (1958). *Statistical Designs of the Health Household Interview Survey*. Washington: Public Health Series, 584-A2, 15-18.
- ORGANISATION DES NATIONS UNIES (1981). Principes et recommandations concernant les recensements de la population et de l'habitation, série *Études statistiques*. Série M, n° 67 (F.80XVII.8). PLATEK, R., RAO, J.N.K., SÄRNDAAL, C.E., et SINGH, M.P. (1987). *Small Area Statistics*. New York: Wiley-Interscience.
- PATTERSON, H.O. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, séries B*, 16, 140-149.
- PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas. *International Statistical Review*, 48, 3-18.
- REDFERN, P. (1987). *A Study of the Future of the Census of Population: Alternative Approaches*. Luxembourg: Statistical Office of European Commission No. ISBN 92-825-7429-6.
- REDFERN, P. (1989). Population registers: Some administrative and statistical pros and cons. *Journal of the Royal Statistical Society, séries A*, 152, 1-41.
- STATE STATISTICAL BUREAU (1987). *The 1987 Nationwide One-percent Population Sample Survey*. Beijing: State Statistical Bureau.
- TREWIN, D. (1987). Estimation of trends and time series models from continuing surveys. *Bulletin of the International Statistical Institute*, 46^e session.

Une taille moins élevée que la normale pour p est tout de même satisfaisante pour estimer des niveaux pour la période courante et des variations nettes avec des estimations pondérées; les optimums sont très peu sensibles et les panels p qui constituent 25 à 50% de l'échantillon global représentent tous ce qu'il y a de mieux ou presque; p peut aussi avoir une taille inférieure à la normale lorsque l'accent est mis sur les échantillons non chevauchants $a - b - c - d$ dans le cas d'aggrégations.

6. CONCLUSIONS ET QUESTIONS

La notion d'échantillons agrégés est à la base de quatre nouvelles méthodes que nous venons d'exposer: échantillon avec renouvellement complet, recensement par étapes, aggrégation asymétrique et plan à panel fractionné. Les échantillons avec renouvellement complet existent déjà mais les trois autres n'ont pas encore été mis en pratique. Entretemps, il serait bon d'approfondir la méthodologie de manière à pouvoir définir les paramètres de faisabilité. Cependant, les principales formes d'application de ces méthodes doivent se trouver dans des situations concrètes et non dans des généralités théoriques. Pour chaque situation, il faudra définir des facteurs de coût, des variances, des biais et des possibilités propres; il faudra trouver une façon de sensibiliser le public aux nouvelles méthodes. Pour illustrer ces propos, nous ne pouvons que soulever quelques questions qui s'ajoutent à celles soulevées implicitement ou explicitement dans les sections précédentes.

1. Quelles sortes de moyennes mobiles peuvent s'avérer les plus utiles pour les échantillons avec renouvellement complet et les recensements par étapes? Dans le cas d'aggrégats nationaux, on peut attribuer tout le poids au dernier mois (ou trimestre ou année). Mais pour ce qui est des petites régions, on peut agréger les données sur une période de dix ans . . . avec des poids égaux ou croissants? Les estimateurs "de réduction" (James-Stein) sont-ils utiles dans les circonstances?

2. Quel effet peut avoir sur les agrégats un changement de population, de méthode, de variable?

3. L'aggrégation asymétrique soulève des questions semblables. Les dernières estimations mensuelles (A) devraient-elles être imprimées avec les estimations agrégées (C)? Il faut trouver des méthodes pour faire correspondre les fréquences de case et les fréquences marginales.

4. En ce qui concerne le plan à panel fractionné, quelle devrait être la taille relative de l'échantillon chevauchant (p)? Peut-il s'agir d'un panel ou simplement de segments chevauchants? Ou doit-il s'agir des deux et si oui, est-ce possible? De quelle façon dépend-il des corrélations pour diverses variables? Comment concilier les quatre grands buts des enquêtes à passages répétés?

Il y aurait d'autres questions tout aussi intéressantes mais nous devons nous arrêter ici pour l'instant.

BIBLIOGRAPHIE

- ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-75.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304.
- KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society* (A), 149, 149-52.
- KISH L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH L. (1979). Samples and censuses. *International Statistical Review*, 47, 99-109.

grands ou des échantillons avec renouvellement complet. Une fois combinés, ces échantillons représentent la meilleure série d'échantillons à chevauchement partiel qui puisse exister pour établir des estimations de niveau et de la variation nette; ils peuvent donc remplacer les échantillons avec renouvellement ordinaires. Cette forme de combinaison est propre au PPF; de plus, ce type de plan permet d'obtenir un échantillon souple qui peut atteindre de grandes dimensions et qui comporte une partie non chevauchante pouvant servir à l'agrégation d'échantillons.

b. Les plans respectifs de p et de $a - b - c$ peuvent être différents et indépendants l'un de l'autre, chacun étant "optimalisé" en vue de ses objectifs propres. Néanmoins, on doit les combiner pour obtenir des estimations conjointes des niveaux et des variations nettes et pour cela, il faut que les populations concernées et les mesures utilisées soient assez semblables.

c. Le PPF présente des avantages considérables par rapport aux plans avec renouvellement classiques du fait qu'il prévoit des échantillons chevauchants pour *toutes* les périodes. Ceci représente un avantage indéniable en ce qui concerne le calcul des variations nettes puisqu'il joue dans toutes les comparaisons voulues. Cet avantage sert aussi pour le calcul d'estimations de niveau car il peut y avoir des différences de corrélation entre les variables.

d. Un avantage important du PPF par rapport aux échantillons à chevauchement partiel actuellement utilisés est la possibilité d'introduire des panels p formés des éléments nécessaires pour mesurer des variations au niveau individuel (variations brutes ou micro-variations). Toutefois, on peut satisfaire les autres caractéristiques au moyen d'échantillons chevauchants p' tirés de segments aréolaires, comme c'est le cas actuellement. De plus, il est possible d'incorporer au plan du panel p ou de l'échantillon chevauchant p' un mécanisme de renouvellement modéré de manière à conserver la plupart des gains issus des covariances et des données du panel. Il y aurait peut-être lieu de prévoir une formule d'alternance afin d'éviter l'effritement du panel. Dans plusieurs enquêtes, on utilise à la fois l'échantillon chevauchant p' et l'échantillon constant p pour suivre le plus grand nombre possible de personnes ayant déménagé. La plupart des ménages font partie des deux échantillons. Le coût additionnel pour l'échantillon constant (ou panel) dépend de la proportion des personnes ayant déménagé et des frais engagés pour les contacter (Kish 1987, 6.2, 6.4).

e. Les avantages et les inconvénients de l'interview de panels soulèvent des questions délicates et font l'objet d'ouvrages nombreux et variés qui renferment parfois des résultats contradictoires (Kish 1987, sections 6.4, 6.5). Il faut établir le nombre et la fréquence des réinterviews qu'il est possible et souhaitable de faire et qui peuvent donner des résultats fiables. Un autre avantage du PPF est qu'il permet de considérer séparément le panel p de manière à pouvoir comparer ses données agrégées à celles des échantillons non chevauchants et à déceler le cas échéant des biais de panel et même à corriger ces biais.

Un autre aspect utile du PPF est la possibilité de prélever des unités de sondage dans le panel à l'aide de différents taux ("optimums") selon que ces unités se trouvent ou non dans l'échantillon non chevauchant.

f. Les échantillons non chevauchants $a - b - c - d$ n'ont pas nécessairement toujours la même taille; cette souplesse caractéristique du PPF, qui tranche avec la rigidité des plans avec renouvellement classiques, peut être utile dans les cas où il faut accroître les échantillons pour réduire les dépenses. De telles mesures ont pour effet de créer des problèmes de pondération pour les données agrégées mais ces problèmes sont solubles.

g. Le rapport entre la taille du panel p et celle des échantillons non chevauchants $a - b - c - d$ dépend des diverses possibilités et de l'étude de coûts et de besoins (section 6). Lorsqu'il s'agit de mesurer des variations au niveau individuel, la taille de p doit être plus élevée que celle de $a - b - c - d$ tandis que c'est l'inverse lorsqu'il s'agit de faire des agrégations. Le fait que p soit généralement plus grand que les échantillons non chevauchants s'explique probablement par le coût relativement peu élevé (sur le terrain) des réinterviews téléphoniques.

période d'un an ou plus pour des variables comme le revenu, etc. D'une part, il y a les recensements, où on sacrifie l'actualité des données et où l'on préfère le dénombrement complet à l'échantillonnage (et aux économies qui en découlent); d'autre part, il y a les enquêtes mensuelles sur la population active et la santé et sur bien d'autres sujets, où l'on insiste sur l'actualité des données et la réduction des coûts au détriment du détail géographique. Les différences de population entre les provinces imposent contraintes à l'actualité des données et à la taille des échantillons. Souvent, on définit un taux de sondage plus élevé pour les provinces moins peuplées mais ce taux, "optimal", aura malheureusement pour effet d'accroître la variance de façon générale et à l'intérieur des "classes de recoupement" pour plusieurs provinces (âge, sexe, etc.) (Kish 1988, section 5; Trewin 1987). Par conséquent, des rapports de tailles de 50 contre 1 ou de 100 contre 1.

A cause de ces difficultés, les tableaux d'enquêtes mensuelles contiennent souvent des cases qui ont une fréquence insuffisante pour les petites provinces. Deux possibilités s'offrent à nous dans les circonstances: A) publier les mêmes données pour les cases à faible fréquence et les cases à fréquence élevée et mettre en garde le lecteur (utilisateur, consommateur) en insérant un avertissement dans une annexe sur les erreurs d'échantillonnage, ou B) ne pas publier les données telles quelles mais supprimer les cases à faible fréquence après s'être fixé certaines limites. Le lecteur peut être orienté vers d'autres publications qui renferment des données agrégées (trimestrielles, annuelles).

L'agrégation asymétrique est une solution intermédiaire entre les publications symétriques (A) et la suppression asymétrique (B). C. L'agrégation asymétrique consiste à produire pour les cases à faible fréquence certaines données périodiques agrégées. La période visée par l'agrégation peut varier, par exemple agré-gation trimestrielle pour les cases à faible fréquence et annuelle pour les cases à très faible fréquence, contrairement à des données mensuelles pour les cases à fréquence élevée. Le lecteur pourrait être averti par un moyen quelconque (astérisque, italique ou autre symbole) du choix qui s'offre à lui; il pourrait alors choisir entre C (agrégation) et B (suppression). AC. Cette procédure permettrait au lecteur de choisir entre A, B et C en publiant à la fois les données mensuelles courantes (A) et les données agrégées selon la formule C. L'inconvénient des formules B et C est que la somme des fréquences de case ne correspond pas à la fréquence marginale, ce qui n'est pas le cas des formules A et AC. On peut remédier à cette difficulté à l'aide d'une méthode itérative.

5. PLANS PANEL FRACTIONNÉ (PPF) USAGES MULTIPLES

Afin d'assurer les sommes nécessaires pour la réalisation d'un recensement par étapes et la création d'un échantillon avec renouvellement complet, il faudrait trouver un moyen d'intégrer l'un et l'autre aux enquêtes périodiques qui existent déjà dans de nombreux pays. Ces enquêtes sont soit mensuelles, soit trimestrielles (parfois annuelles ou hebdomadaires). Les échantillons qui y sont utilisés sont essentiellement des échantillons à chevauchement partiel, qui visent à procurer des estimations de niveau (pour la période courante) et de la variation nette plus précises. Toutefois, ces échantillons ne sont pas conçus pour des enquêtes qui utilisent des échantillons agrégés ni pour des enquêtes par panel fondées sur des échantillons chevauchants. Nous proposons plutôt le plan à panel fractionné (PPF) pour concilier les usages ci-dessus; en outre, ce plan comporte certains avantages secondaires (Kish 1987, 6.5). a. Le PPF est essentiellement la combinaison de deux échantillons périodiques distincts: un échantillon constant p est jumelé à une série parallèle d'échantillons non chevauchants $a - b - c - d$, etc.; on désigne le nouvel échantillon par $pa - pb - pc - pd$, etc. L'échantillon constant permet de mesurer les variations au niveau individuel (micro-données) tandis que les échantillons non chevauchants peuvent être groupés pour former des échantillons plus

Le plus vieux cas d'agrégation de données que nous avons pu trouver a trait à un échantillon prélevé en Californie en 1952 (Mooney 1956). 'Les échantillons ont été prélevés de manière à obtenir un taux de sondage uniforme global de 1 sur 385, des fins de recensement, on a divisé l'échantillon en 52 groupes égaux et on a recensé chacun de ces sous-échantillons à tour de rôle durant l'année d'enquête à raison d'un par semaine. Par conséquent, chaque recensement hebdomadaire reposait sur un échantillon de 1 sur 20,020' (traduction). Pour de plus petits Etats (plus petites populations) ou de plus grands échantillons, on peut penser à des échantillons hebdomadaires de 1/520 et à un échantillon cumulatif au bout des 520 semaines qui forment la période inter-censitaire. Ce genre d'échantillon a vraisemblablement été conçu pour de petites populations.

Les exemples ci-dessus ont trait à des échantillons périodiques non chevauchants. On s'est servi de données agrégées provenant d'échantillons à chevauchement partiel mais la 'taille effective de l'échantillon' a été diminuée du nombre d'unités incluses dans la portion commune (Ericksen 1974). De plus, cet article porte sur l'agrégation de données relatives à des individus; cependant, les enquêtes à passages répétés peuvent aussi servir à combiner des données relatives aux individus (Kish 1987, 6.6).

4. AGREGATIONS ASYMÉTRIQUES

Le titre de cette section désigne une méthode d'agrégation qui est proposée pour les cas où la taille des sous-populations 'naturelles' varie beaucoup. Par exemple, nous avons observé ces dernières années des rapports de tailles de 50 contre 1 et même de 100 contre 1 parmi les provinces (ou les Etats) du Canada, des Etats-Unis, de l'Australie et de la Chine; cet intervalle de rapports de tailles est le même pour la plupart des pays. Cette inégalité vient du fait que les unités administratives ont à peu près toutes la même superficie mais correspondent à des territoires très inégalement peuplés. On observe aussi cette inégalité pour des districts, des comtés, etc. dans la plupart des provinces ou Etats. Le phénomène existe aussi pour d'autres unités et organisations sociales telles que les entreprises, les hôpitaux, les universités. Il y a toutefois des unités qui échappent à ce phénomène: les unités militaires, les districts de recensement et les territoires d'écoles primaires ont à peu près tous la même superficie.

Pour de nombreuses autres distributions de fréquence, on crée des classes à peu près égales en agrégeant des données, selon les méthodes généralement reconnues, en fonction d'échelles approximativement logarithmiques; par exemple, le revenu, la taille d'une agglomération, etc. sont des variables qui sont souvent divisées en classes (10-25, 25-50, 50-100, 100-250, 250-500, 500-1000, 1000-2500, etc.). C'est là une méthode d'agrégation commode par laquelle on crée des cases à fréquences à peu près égales sur une échelle approximativement logarithmique; ce genre de cases sont acceptées de tous depuis longtemps bien qu'elles soient fortement asymétriques. Il convient aussi de souligner que les cases des tableaux de données d'échantillon renferment des fréquences qui sont *généralement agrégées dans l'espace et le temps*. Par exemple, les enquêtes mensuelles sur la population active montrent souvent des données agrégées pour une période d'un mois (ou d'une semaine, laquelle est représentative du mois) ou pour des provinces (à partir d'un échantillon d'unités de sondage). Il y a aussi des données agrégées en fonction du trimestre ou de l'année et en fonction du territoire national. La période visée par l'agrégation est assujettie à trois contraintes simultanées: l'étendue de la période de référence, qui peut être relativement variable; les domaines de sous-populations, qui peuvent être rigides, comme les provinces; et la taille de l'échantillon, exprimée par les unités de sondage et les composantes de la variance. D'autres variables, comme les facteurs de coût et le degré de précision requis, sont généralement exprimées par les trois paramètres fondamentaux de la taille des cases.

Les recensements décennaux de la population représentent un cas extrême par leur souci du détail géographique; chaque personne est associée à une adresse en particulier à la date de référence (le 1er avril aux E.-U.). Néanmoins, il est possible de faire des agrégations sur une

dans le temps sont plus difficiles à faire accepter que les plans qui prévoient un calcul de variations dans l'espace. Les variations peuvent être fortes et parfois appréciables, mais elles sont le plus souvent accidentelles. Néanmoins, nous avons appris à accepter des échantillons, des moyennes et des agrégations de ces deux catégories d'éléments sous forme d'aggrégats et de moyennes de population (moyennes et aggrégats nationaux).

Il y a moyen de briser la résistance que l'on oppose encore à l'idée de recensement par étapes et d'échantillon avec renouvellement complet à l'aide d'arguments théoriques et pragmatiques. Les arguments théoriques ont été évoqués plus haut ainsi que dans d'autres études portant sur des solutions de rechange (Kish 1987, 6.1B). Quant aux arguments pragmatiques et empiriques, nous pouvons les soutenir en parlant de plusieurs types d'utilisations qui sont reconnus comme courants et efficaces. Les échantillons périodiques qui servent à mesurer des variations et à recueillir des données pour la période courante peuvent aussi servir à calculer des aggrégats pour des régions et des domaines. De plus, en faisant la moyenne pour une année ou plus, on se trouve à lisser les variations temporelles (saisonnnières, cycliques ou accidentelles) à l'aide de moyennes mobiles.

Données rétrospectives. Le nombre d'enfants nés de femmes qui ont franchi la période de fécondité de 30 ans peut représenter un cas extrême pour des données rétrospectives. Cependant, il existe d'autres données individuelles que l'on peut agréger pour toute la durée de vie, par exemple des données sur les maladies graves, le niveau d'instruction, etc. Les données d'interview agréées pour une période d'un an portent sur la production agricole, les antécédents professionnels, le revenu, les achats de maisons et d'automobiles. Toutes ces données ont évidemment des lacunes, qui varient selon les variables, les répondants, les méthodes, etc. Même les données agréées pour une période d'une semaine ou d'une journée (comme les achats de pain ou de cigarettes) renferment des erreurs. Les *sondages échelonnés* servent à agréger des données à court terme; par exemple, le nombre de naissances enregistrées au cours d'une année a été établi à l'aide de 12 échantillons mensuels prélevés dans l'année.

L'agrégation de données rares provenant d'enquêtes à passages répétées a souvent été utilisée pour résoudre ces problèmes délicats et coûteux. Le sujet a été traité et illustré dans des publications portant sur les questions peu courantes (Kish 1965, 11.4; Kalton et Anderson 1986). Les données pour petits domaines peuvent aussi tirer profit de l'agrégation et une année de naissance peut représenter un petit domaine, qui est constitué de "classes de recoupement". Or, les unités géographiques et administratives sont des "domaines propres"; les échantillons périodiques ne conviennent pas dans ce cas puisque les domaines de ce genre exigent un recensement par étapes ou un échantillon avec renouvellement complet.

Aggrégation de données tirées d'échantillons périodiques. La Health Household Survey (NCHS, 1958), que nous avons décrite plus haut, est probablement l'exemple le plus connu d'une enquête où on agrège pour un an les données d'échantillons hebdomadaires d'environ 1,000 ménages tirés de segments aérolaires non chevauchants. Il s'agit d'une enquête à *objectifs multiples* (comme la plupart des enquêtes à passages répétées), notamment l'agrégation de données portant sur des maladies rares, le calcul d'estimations (de niveau) pour la période courante et le calcul de variations nettes. Cette enquête produit des estimations pour de plus grands domaines ainsi que des estimations nationales sur les maladies les plus courantes. Si on utilisait pour cette enquête un échantillon avec renouvellement complet, en augmentant l'étendue des échantillons annuels, on accroîtrait le coût des opérations sur le terrain, particulièrement dans la portion du territoire (environ 30 pour cent de celui-ci seulement) où les u.p.é. sont des comités (unités non auto-représentatives).

Une enquête de circulation est un exemple intéressant d'agrégation de données parce que la population est très mobile à l'intérieur de la base de sondage qui est formée d'unités d'échantillonage du type "lieux heure" (Kish, Lovejoy et Rackow 1961). Le concept général de l'enquête peut s'appliquer aux nomades et à d'autres populations mobiles. Il peut aussi s'appliquer à des populations générales moins mobiles pour une période plus longue, comme celle qui sépare deux recensements décennaux.

et les inconvénients des trois catégories de données sont complémentaires; ce serait donc une bonne chose de combiner les avantages des trois catégories. C'est ce que visent les nombreuses méthodes d'estimation pour petites régions: produire des estimations actuelles, précises et pertinentes pour les petites régions et d'autres petits domaines.

Ces méthodes servent actuellement à estimer la population de petites régions entre deux recensements décennaux pour compenser l'effet de l'obsolescence des données du recensement décennal; c'est pourquoi les estimations obtenues par ces méthodes sont parfois appelées *estimations postcensitaires*. Les méthodes d'estimation pour petites régions ont des usages de plus en plus nombreux; par exemple, on a proposé de les utiliser pour compenser les biais dus au sous-dénombrement. Toutefois, elles sont l'une et l'autre une combinaison de recensement, d'enquête par sondage et de fichier administratif. Il ne faudrait donc pas les voir immédiatement comme un substitut des recensements. Néanmoins, il est permis de se demander si un recensement par étapes ne donnerait pas de meilleurs résultats qu'un recensement décennal dans cette combinaison. Nous ne pouvons donner de réponse catégorique mais nous croyons que le recensement par étapes serait supérieur dans la plupart des cas à cause de la valeur des composantes de la variance. Cependant, des études théoriques aussi bien qu'empiriques seront nécessaires pour résoudre cette question et bien d'autres que nous nous posons.

Nous devons aussi arrêter aux *échantillons à chevauchement partiel* des plans à usages multiples car de nombreux pays les utilisent à plusieurs fins et leur affectent des fonds à même le budget prévu pour l'établissement de statistiques nationales. Les plans de sondage à usages multiples permettent le plus souvent de recueillir des données sur la population active et d'autres données utiles. Ils varient d'un pays à l'autre dans le choix de leurs paramètres mais, partout dans le monde, ils partagent plusieurs caractéristiques fondamentales avec les plans utilisés aux Etats-Unis et au Canada. Ils produisent des échantillons périodiques chevauchants, dont la portion commune est fixe pour une période déterminée (les trois paramètres varient toute-fois selon les pays). Ils utilisent des segments aréolaires comme bases de sondage, mais non des panels de ménages (on ne se préoccupe pas des personnes ayant déménagé). Le taux de chevauchement est généralement élevé et cela s'explique habituellement par le fait qu'un taux élevé favorise une réduction de la variance attribuable à la corrélation positive qui existe entre les éléments de la portion constante. Un avantage encore plus précieux des échantillons chevauchants est peut-être la réduction des frais d'interview dans les contacts ultérieurs, particulièrement dans le cas d'appels téléphoniques faits à la suite de visites à domicile. Ces "plans avec renouvellement" ont fait l'objet de très nombreuses études dans le domaine statistique et représentent une innovation importante (attribuée à H.D. Patterson 1950 et R.J. Jessen 1942). Ils sont conçus pour mesurer des variations nettes et établir des estimations (de niveau) de la période courante mais non pour agréger des données. Néanmoins, si nous avions un taux de chevauchement beaucoup moins élevé que le taux couramment utilisé (par exemple, moins de 30% comparativement à plus de 70%), la variance (par ménage) n'augmenterait pas de beaucoup. Cela est particulièrement vrai pour de nombreuses variables, comme le fait d'être en chômage, pour lesquelles il y a peu de corrélation entre les périodes. De plus, il y a d'autres façons de modifier les échantillons chevauchants (section 5). Par conséquent, il serait possible de combiner ces plans de sondage avec les agrégations requises pour les recensements par étapes et les échantillons avec renouvellement complet.

3. AGREGATION DANS LE TEMPS ET L'ESPACE

Les variations observées pour les populations et les variables correspondantes sont souvent de trois types: les tendances séculaires, qui sont plus ou moins lisses et monotones, comme la "croissance"; les variations cycliques ou périodiques, comme les variations saisonnières; et les variations accidentelles, qui sont difficiles à décrire et sont souvent considérées comme "aléatoires". Les plans qui prévoient une agrégation, un calcul de moyennes et un échantillonnage

les opérations de traitement à une organisation beaucoup plus modeste et mieux préparée ainsi qu'à du personnel plus expérimenté (. . .) Le recensement par étapes n'attirerait pas autant d'attention qu'un recensement ordinaire. Bien que cela pourrait contribuer à réduire les risques de mécontentement dans le public, on obtiendrait un taux de couverture moins élevé que la normale (. . .) (La méthode utilisée) compliquerait l'interprétation des résultats du recensement et plus spécialement les comparaisons entre régions. La simultanéité des opérations sur le territoire national, qui est une des qualités fondamentales du recensement classique, n'existerait plus. Le recensement par étapes est un concept qui est encore tout théorique" (traduction) (2.13).

La majorité des pays auront probablement encore besoin des recensements au 21^{ème} siècle. En 1990, les recensements sont devenus graduellement au profit des registres de population dans les pays nordiques tandis que leur usage est encore inconnu dans certains pays du Tiers-Monde. On s'oppose même à leur utilisation dans une très faible minorité de pays. Néanmoins, la majorité des Etats ont besoin de recensements et s'approprient justement à réaliser le leur en 1990. L'instar de la locomotive à vapeur, qui date à peu près de la même époque, les recensements ont été une invention précieuse mais il se peut que, comme la locomotive, ils soient remplacés tôt ou tard par d'autres formes de dénombrement.

On a déjà proposé des *recensements annuels et quinquennaux* et on relève des cas d'utilisation de recensement quinquennal dans quelques pays, dont le Canada et la Turquie. Cependant, nous avons le sentiment que la formule du recensement quinquennal n'est pas près d'être acceptée à cause de son coût exorbitant: dans deux pays qui ont eu recours au recensement quinquennal, un échantillon de dix pour cent a coûté la moitié de ce que coûte un dénombrement complet. Par ailleurs, les recensements annuels (qui sont en réalité des recensements par sondage - à 5 ou 10%) ne produiraient pas de données assez détaillées sur le plan géographique. Le "micro-recensement" (à 1%) de l'Allemagne de l'Ouest produit des données d'échantillon annuelles. En 1987, la Chine a réalisé un recensement à 1%; ses échantillons annuels de 1/2000 (ce qui équivaut à environ 500,000 personnes) servent à recueillir uniquement des données sur la fécondité (SSB 1987; Kish 1989). En conclusion, les recensements quinquennaux ne sont pas assez fréquents et la formule des *recensements annuels* serait trop coûteuse.

De nombreux pays possèdent des *fichiers administratifs* riches en données de toutes sortes et le nombre de ces fichiers est censé s'accroître dans l'avenir. On trouve d'excellents *registres de population* dans des pays comme la Suède, la Norvège, le Danemark et la Finlande et probablement dans d'autres pays d'Europe septentrionale. La richesse de ces fichiers est principalement attribuable à la coopération, à la motivation et au degré d'instruction élevé des répondants dans ces pays; dans certaines occasions, on utilise les données des registres de population au lieu de faire un recensement. Dans d'autres circonstances, la qualité de ces registres laisse fortement à désirer. Nous pouvons nous attendre à une amélioration de la qualité de ces registres et à un accroissement de leur utilisation dans un avenir plus ou moins rapproché. Cependant, il serait bien étonnant qu'ils remplacent les recensements même dans des pays développés comme les E.-U. et le Canada et leur utilisation prochaine dans les pays en voie de développement est encore plus hypothétique (Redfern 1989).

Et même si la qualité et le champ de ces registres devenaient acceptables, ceux-ci ne contiendraient jamais d'autre chose qu'un minimum de variables démographiques: chiffres de population, âge, sexe et quelques autres éléments. Par conséquent, les registres de population ne pourront répondre aux exigences de la société moderne en ce qui concerne les sources de données. Ils ne seront jamais plus qu'une source de variables auxiliaires.

Les estimateurs synthétiques, les estimateurs par quotient et les estimateurs de la méthode itérative du quotient sont trois types d'estimateurs qui sont de plus en plus utilisés pour l'établissement d'estimations régionales (Platek et coll. 1987; Purcell et Kish 1980). Les données de recensement sont souvent pémées, celles des fichiers administratifs sont incomplètes et les données d'échantillon ne sont pas suffisamment détaillées pour les petites régions. Les avantages

2. AUTRES MÉTHODES DE RECENSEMENT

Le recensement par étapes serait une opération coûteuse mais on pourrait en justifier l'utilisation en invoquant les lacunes reconnues des recensements décennaux, qui sont très à la mode actuellement, de même que des enquêtes par sondage et des fichiers administratifs, que l'on propose parfois comme solutions de rechange. Le principal motif des recensements est la nécessité de recueillir des données détaillées, spécialement pour les petites régions, et la principale faiblesse des recensements décennaux est la trop grande période qui s'écoule entre deux recensements, ce qui favorise l'obsolescence des données recueillies, de même que leur coût total très élevé, qui empêche la réalisation de recensements plus fréquents. Les enquêtes par sondage présentent de nombreux avantages pour ce qui a trait aux statistiques nationales et aux grandes régions mais elles ne produisent pas de données assez détaillées sur le plan géographique et sur d'autres plans aussi. Les fichiers administratifs de qualité sont rares et les données qu'ils contiennent se limitent le plus souvent à des variables démographiques.

L'usage des recensements décennaux de la population, du logement, de l'agriculture, de l'industrie, etc. s'est implanté dans la plupart des pays au cours des deux derniers siècles, et plus particulièrement au cours des deux dernières générations grâce à l'action du Bureau de statistique des Nations-Unies. Outre qu'ils produisent des données détaillées pour les petits domaines, les recensements procurent souvent un meilleur taux de couverture que les enquêtes par sondage à cause de la forte publicité dont ils font l'objet et des "activités" nationales auxquelles ils donnent lieu; le recensement de 1982 en Chine illustre bien cela (Kish 1978, 1989). Par ailleurs, les coûts unitaires d'un recensement seront beaucoup moins élevés que ceux d'une enquête par sondage à cause de la concentration des activités mais les coûts totaux, beaucoup plus élevés à cause de l'envergure des opérations de recensement. Évalués à 2,6 milliards de dollars, les recensements qui seront réalisés aux E.-U. en 1990 reviendront à 10 \$ par habitant ou à 30 \$ par ménage. Ce coût, compris entre une demi-fois et une fois le salaire horaire médian (calculé sur une période de dix ans), est comparable à celui enregistré dans d'autres pays, quoique le nombre et la complexité des variables de recensement soient en l'occurrence des éléments importants de ce coût. Le recensement par étapes serait probablement la solution indiquée pour les cas où on compte un grand nombre de variables passablement complexes. Au Canada, 260 échantillons hebdomadaires de 32,000 ménages chacun seraient nécessaires pour couvrir toute la population. Aux États-Unis, l'agrégation de 520 échantillons hebdomadaires de 160,000 ménages chacun donnerait au bout de dix ans 80 millions de ménages; à l'heure actuelle, l'échantillon de la Current Population Survey (CPS) compte 100,000 ménages, y compris les ménages supplémentaires par État.

Nous ne pouvons faire ici une comparaison détaillée des recensements décennaux et des recensements par étapes; cependant, il faut souligner la question de l'actualité des données car c'est là l'élément fondamental de la comparaison. En règle générale, le délai de publication des données d'un recensement décennal varie de un à quatre ans et ces données peuvent ensuite servir pendant plus de quatorze ans. Malgré l'existence d'ordinateurs plus rapides de nos jours, le délai de publication est toujours plus long pour des statistiques sociales complexes que pour de simples chiffres de population; de plus, les données d'un recensement décennal deviennent de plus en plus rapidement aujourd'hui à cause de la plus grande mobilité de la population. Les biais dus à l'obsolescence seront monotones, sinon linéaires. Leur grandeur variera selon la variable, la population, etc., mais nous croyons qu'ils seront considérables et même peut-être supérieurs au fameux biais dû au sous-dénombrement (Kish 1981, 1979).

Compte tenu de l'obsolescence rapide des données de recensements décennaux, nous devrions nous attacher à trouver des solutions de rechange, comme le fait Redfern dans son étude intitulée *A Study on the Future of the Census of Population: Alternative Approaches* (Redfern 1987). "Un inconvénient majeur des recensements", dit-il, "est qu'ils reviennent relativement peu souvent" (traduction) (2.3). Au sujet du recensement par étapes, il écrit: "L'avantage de cette proposition est qu'elle permettrait de confier les opérations sur le terrain de même que

pour obtenir les chiffres (pondérés) d'un recensement décennal ou ceux d'un sondage annuel de la population à dix pour cent.

Dans la Health Interview Survey du National Center for Health Statistics (1958), on agrège les données de 52 échantillons hebdomadaires comptant environ 1 000 ménages chacun. Le taux de sondage hebdomadaire est environ $f = 1/80,000$; au bout de dix ans, les échantillons périodiques non chevauchants représentent donc un taux de sondage cumulatif de $520/80,000$. Cependant, ces échantillons sont toujours limités à la même série d'u.p.é. par souci d'économie surtout, mais aussi pour obtenir de meilleures estimations de la variation nette ainsi que des estimations de la période courante. En revanche, les *échantillons avec renouvellement complet* sont peut-être mieux conçus pour maximiser (accroître) le champ (la représentativité) des échantillons prélevés successivement dans la population nationale (la population en général). Les termes entre parenthèses indiquent que les échantillons avec renouvellement complet sont un cas particulier des *échantillons périodiques cummulatifs* et qu'il n'est pas essentiel de définir clairement la limite entre les premiers et les seconds.

Dans le cas des échantillons périodiques chevauchants, la sélection des unités est soumise à des exigences tout à fait contraires à celles qui déterminent le choix des objectifs et du contenu des interviews (observations, variables). Les questionnaires doivent être aussi uniformes que possible pour tous les passages d'une enquête pour que les données agrégées soient significatives. Si on ajoutait des questions sur des sujets nouveaux mais en utilisant des échantillons périodiques formés toujours des mêmes éléments, on accroîtrait la diversité des renseignements de l'enquête mais non la taille de l'échantillon pour les données d'enquête. Dans la plupart des enquêtes à passages répétés, on recueille à chaque fois des données sur les mêmes variables; il arrive cependant que des enquêtes renferment des sujets additionnels. Néanmoins, un changement de méthode, de question ou de variable peut causer de sérieux problèmes. La meilleure façon, peut-être, d'opérer de tels changements serait de procéder par "raccourcissement", en utilisant à la fois l'ancienne et la nouvelle méthode pour analyser les différences. Les problèmes que nous évoquons ici sont essentiellement comparables à ceux que l'on retrouve lorsqu'on évalue les différences entre les passages d'une enquête sauf qu'ils paraissent plus inédits. Dans la section 6, nous maintenons que ce genre de problèmes ne peuvent être résolus que par une formule adaptée à chaque situation.

Par ailleurs, le fait de toujours recueillir des données auprès des mêmes éléments (personnes, ménages) ne contribue pas à accroître proportionnellement la taille de l'échantillon (base) et les panels formés des mêmes éléments n'ajoutent rien aux échantillons avec renouvellement complet. Dans de nombreuses enquêtes à passages répétés (par ex.: enquêtes sur la population active du Canada, des E.-U., etc.), des fractions de segments (groupes de dernier niveau) peuvent se chevaucher plus ou moins fortement et elles ne contribuent pas vraiment à accroître la taille de l'échantillon. Même dans les enquêtes où les segments ne se chevauchent pas (par ex.: les Health Interview Surveys du NCHS (1985)), ceux-ci sont toujours limités aux mêmes unités du premier degré (et du second degré?); les corrélations positives (effets de grappe) qui existent dans ces segments ont tendance à réduire la taille "effective" de l'échantillon pour des données globales. De plus, comme les échantillons périodiques de ces enquêtes sont restreints à un échantillon d'unités primaires, ils ne répondent pas au critère qui s'applique aux échantillons avec renouvellement complet, à savoir la couverture de la population entière (nationale?). Quelques remarques additionnelles peuvent servir à élargir notre cadre de référence. 1) Dans des analyses de ce genre, on suppose souvent un échantillonnage aréolaire (base aréolaire); toutefois, la discussion peut s'étendre à d'autres genres de bases de sondage. 2) Il est souvent question aussi de probabilités de sélection égales, mais on peut tout aussi bien opter pour un échantillonnage à intervalles irréguliers dans le temps. 3) De même, on peut aussi penser à un échantillonnage avec probabilités de sélection inégales. 4) Une agrégation (ou accumulation) de données d'échantillon pour la période complète (année ou décennie) est la chose qui nous vient le plus facilement à l'esprit mais nous pouvons envisager aussi un échantillonnage systématique à l'intérieur de la période visée; par exemple, les enquêtes sur la population active portent sur une semaine particulière à chaque mois de l'année (voir figure 1.1).

Recensement par étapes et échantillons avec renouvellement complet

LESLIE KISH¹

RÉSUMÉ

Un recensement par étapes consiste en F échantillons périodiques non chevauchants qui représentent chacun un taux de sondage de $1/F$ et qui sont conçus de telle manière qu'en faisant la somme des données des F échantillons, on obtient un dénombrement complet de la population du territoire visé avec $F/F = 1$. En agrégeant les données de k échantillons ($k < F$), on obtient un échantillon plus global représentant un taux de sondage k/F , qui répond à des besoins plus ponctuels (recensements annuels ou quinquennaux). Dans le cas de populations naturellement mobiles, les bases de sondage aréolaires couvriront le territoire national. Ces méthodes pourront souvent être préférées à d'autres méthodes de recensement que nous traitons aussi dans cet article. L'agrégregation asymétrique est une méthode recommandée pour résoudre les problèmes liés à l'existence de cases à faible fréquence pour les domaines aréolaires (provinces, régions, États), problèmes du reste communs à la plupart des pays et à d'autres unités de population. Les plans à panel fractionné (PPF) sont un autre moyen d'agréger des données d'enquêtes périodiques; le PPF combine un échantillon constant p avec des échantillons non chevauchants $a - b - c - d$ de sorte que $pa - pb - pc - pd$ désignent des échantillons à chevauchement partiel pour des plans à usages multiples.

MOTS CLÉS: Échantillons périodiques; échantillonnage dans le temps; agrégation; plans à panel fractionné; agrégation asymétrique; plans de sondage à usages multiples.

1. INTRODUCTION ET DÉFINITIONS

Dans cet article, nous allons étudier plusieurs façons d'agréger des données d'échantillons périodiques et voir pour quelles raisons cela se fait. Ce sujet a reçu peu d'attention jusqu'à maintenant puisque la plupart des ouvrages portant sur les échantillons périodiques et les échantillons avec renouvellement partiel insistent beaucoup plus sur les variations nettes et les estimations de période courante (estimations "transversales") que sur les agrégations. Nous nous intéressons avant tout au recensement par étapes et aux échantillons avec renouvellement complet et nous tentons ici de définir le recensement par étapes: plan de sondage composé prévoyant F échantillons périodiques distincts (non chevauchants), chacun d'eux étant un échantillon probabiliste de la population avec taux de sondage $f = 1/F$, conçus de telle manière qu'en faisant la somme des données des F périodes, on obtient un dénombrement complet de la population avec $f' = F/F = 1$. L'agrégregation des données de k périodes ($k < F$) devrait se traduire par des échantillons avec renouvellement complet représentant un taux de sondage $f' = k/F$ et qui renferment des données portant sur une à F périodes. Des exemples et des contre-exemples nous permettront de mieux saisir cette définition. De plus, nous examinerons des variantes qui seraient susceptibles de répondre à la définition et aux besoins divergents que peuvent devoir satisfaire les échantillons avec renouvellement complet.

Imaginons un échantillonnage hebdomadaire au niveau national, avec un taux de sondage (avec probabilités égales) de $1/520$, conçu de telle manière qu'au bout de 520 semaines, on a couvert toute la population et la somme des données de chaque échantillon correspond au chiffre réel de la population pondéré sur dix ans. La fin de chaque année de la période en question, on aurait un échantillon national et des échantillons régionaux représentant un taux de sondage de $52/520 = 1/10$. Il suffirait d'agréger les données des échantillons hebdomadaires nationaux

¹ Leslie Kish, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106 E.-U.

organismes statistiques travaillent sur cette question. Avec l'aide de la psychologie, les statisticiens pourraient s'inspirer de ces méthodes pour mettre au point un meilleur modèle du processus de réponse – processus qui est sans doute la moins bien comprise des composantes de l'enquête. Le troisième domaine est le développement de systèmes intégrés d'information statistique alliant des modèles de structures sociales et économiques à des bases de données et permettant de simuler les effets de diverses politiques. De tels systèmes faciliteraient l'utilisation des données d'un organisme pour analyser des politiques et aideraient cet organisme à déceler les lacunes de données dans ses programmes.

Pour faire écho à la conclusion de Bailar, je dirai que les problèmes sont passionnants et qu'il n'en manque pas.

des progrès pour l'avenir; mais des travaux importants y ont déjà été faits. Voici une liste de domaines dans lesquels Statistique Canada a fait des recherches.

- a) **Méthodes d'analyse des données d'enquêtes complexes** Très utiles pour les utilisateurs de la plupart des statistiques gouvernementales, ces méthodes sont conçues pour permettre d'adapter ou pour remplacer les méthodes traditionnelles d'analyse statistique fondées sur l'échantillonnage aléatoire simple. Ce domaine de recherche a suscité l'intérêt des universitaires, qui ont également fait de nombreuses contributions sous ce rapport.
- b) **Couplage d'enregistrements** Cette technique est utilisée pour calculer des valeurs statistiques à partir de données administratives, pour évaluer la qualité des données, dans le cas de couplages de micro-données, et pour la tenue à jour de listes. L'élaboration d'une théorie générale du couplage d'enregistrements a permis de mettre au point des logiciels de couplage. Sur cette question, la plupart des travaux ont été effectués par des organismes statistiques d'État.
- c) **Contrôle et imputation** Largement utilisée dans de nombreuses enquêtes, cette technique a manqué de bases solides jusqu'au moment où l'on a élaboré une théorie, dans les années 70. Depuis, on a mis au point des méthodes et des systèmes d'application générale pour effectuer le contrôle et l'imputation dans diverses enquêtes. Ce sujet a suscité beaucoup d'intérêt et de nouveaux travaux à l'extérieur des organismes statistiques.

- d) **Estimations régionales** Ces dernières années, on s'est de plus en plus intéressé à la production d'estimations pour des régions plus petites que la taille permise par l'estimation directe à partir d'enquêtes par échantillonnage. Les organismes statistiques, avec la participation d'universitaires, ont mis au point diverses méthodes pour résoudre ce problème. Mais l'utilisation de ces méthodes pour la production d'estimations est encore assez limitée, en partie parce que la question demeure de savoir s'il convient pour un organisme gouvernemental de produire des estimations au moyen d'un modèle.
- e) **Utilisation statistique de données administratives** Les dossiers administratifs ont été utilisés comme un moyen parmi d'autres de réduire le coût de la collecte de renseignements. Cette source présente, relativement à la couverture et à la qualité des données, un ensemble de problèmes qui diffèrent de ceux que posent les enquêtes. Bien que les données administratives puissent être utilisées seules pour produire des données statistiques, on les utilisera plus efficacement en les conjuguant à des données d'enquête ou de recensement dans des systèmes d'estimation qui tirent parti des avantages relatifs de chaque type de données. La plupart des travaux dans ce domaine ont été effectués par des organismes statistiques d'État.

5. L'avenir

Considérant l'avenir, Bailar prévoit une utilisation accrue des modèles. Il est presque impossible de douter de la justesse de ce pronostic, à une époque où les organismes statistiques cherchent de plus en plus à tirer le maximum d'information de données existantes et à contourner l'augmentation du coût de la collecte de données. Bailar mentionne en particulier l'intégration des méthodes des séries chronologiques aux méthodes d'estimation, question actuellement étudiée dans plusieurs organismes statistiques. J'ajouterais trois autres domaines dans lesquels on peut espérer des progrès importants à long terme, ces progrès supposant dans chaque cas une interaction entre la statistique et d'autres disciplines.

Le premier domaine est l'application de systèmes experts à certaines activités des organismes statistiques d'État. Pour prendre un exemple dont nous avons déjà parlé, le choix des options ou des modèles qu'il convient d'utiliser pour la désaisonnalisation des séries chronologiques pourrait se prêter à une approche de ce genre. Le deuxième domaine est l'utilisation des méthodes cognitives pour comprendre et améliorer le processus de réponse. Plusieurs

également des modèles qui pourraient remplacer les techniques traditionnelles du modèle X-11 ARMMI. C'est là un domaine de la recherche en statistique auquel s'intéressent les universitaires spécialisés dans les séries chronologiques. Les méthodes de désaisonnalisation ont manifestement des applications qui dépassent l'utilisation que peuvent en faire les organismes statistiques d'État.

Enfin, le domaine dont parle Baillar où les contributions sont le plus récentes est celui de la protection du secret statistique. Il s'agit là d'un problème qui intéresse presque exclusivement les organismes assujettis à des règles de confidentialité interdisant la divulgation de tout renseignement qui pourrait permettre d'identifier une entité enquêtée. Le gros de la recherche dans ce domaine se fait dans les organismes statistiques d'État. Toutefois, les techniques utilisées viennent surtout de l'informatique, de l'analyse numérique et des mathématiques. Ce domaine est assez nouveau et n'a pas encore suscité beaucoup de travaux à l'extérieur des organismes d'État.

Ces exemples montrent que les contributions méthodologiques des organismes statistiques d'État non seulement servent à résoudre les problèmes qui se posent à ces organismes mais peuvent aussi faire faire des progrès importants à la statistique en général. Naturellement, ce ne sont pas toutes ces contributions qui ont un champs d'application très étendu, et certaines pourront rester sans grande utilité à l'extérieur des organismes d'État. Un des défis permanents qui se posent aux statisticiens du secteur public est de susciter l'intérêt de leurs collègues des autres secteurs, en particulier des universitaires, pour les recherches qui se font dans les organismes d'État.

3. Climat favorable à l'innovation

Les innovations sont rarement le fruit du hasard. Il faut un climat qui favorise l'éclosion des idées et le progrès de la recherche. Ce climat n'est pas toujours facile à créer dans une organisation dont la principale tâche est de diffuser régulièrement des données selon un calendrier préétabli. Baillar mentionne trois raisons données par Hansen pour expliquer pourquoi l'enquête par échantillonnage probabiliste a été adoptée assez rapidement au Censur Bureau. Ces trois raisons définissent en substance quelles conditions préalables doivent exister pour qu'un organisme statistique soit un milieu favorable à l'innovation:

- a) le soutien de la direction, c'est-à-dire la volonté d'investir dans la recherche;
- b) la collaboration de clients, c'est-à-dire qu'il faut à la recherche, pour être fructueuse, une application particulière correspondant au problème qui se posait au départ et déterminant un échéancier; le gestionnaire d'un programme de recherche de ce genre doit être quelqu'un qui est à l'aise dans un climat d'expérimentation;

c) des chercheurs compétents, non seulement par leur connaissance d'un domaine en particulier, mais aussi par leur capacité de reconnaître quels problèmes peuvent être généralisés et résolus au moyen des méthodes statistiques.

Ces trois conditions peuvent créer un bon climat pour la recherche, mais d'autres efforts pourront être nécessaires pour faire en sorte que les résultats des travaux soient utilisés efficacement. Cela suppose que le statisticien possède des dons pour persuader et communiquer et que l'organisme où s'effectue la recherche offre le soutien voulu pour la nouvelle méthodologie.

4. Autres contributions

Baillar ne cherchait pas, dans le choix de ses exemples, à couvrir tous les domaines où des contributions ont été faites à la méthodologie statistique. Nous pouvons énumérer ici d'autres domaines de la méthodologie où des organismes d'État ont apporté des contributions importantes. Certains sont mentionnés par Baillar parmi les domaines où elle envisage

Premièrement, considérons les organismes statistiques ailleurs qu'aux États-Unis. Dans la plupart des pays, l'organisme statistique d'État est une institution unique chargée de faire à intervalles réguliers d'importantes enquêtes sur les ménages et les entreprises, d'intégrer les données provenant de diverses sources, de tenir à jour et d'analyser des séries chronologiques et de diffuser de grandes quantités de données dans le public. (Sous ce rapport, les États-Unis sont un cas d'exception du fait qu'ils ont plusieurs grands organismes qui se partagent ces activités selon les domaines concernés.) Dans la plupart des pays, donc, les organismes statistiques doivent regarder à l'étranger pour observer des expériences semblables aux leurs ou pour avoir des échanges entre experts. Le réseau d'interaction entre les organismes statistiques des pays industrialisés est très développé. Les rapports peuvent être bilatéraux ou multilatéraux. L'ancienne et vivante tradition d'échanges d'information et de résultats d'expériences entre Statistique Canada et le U.S. Bureau of the Census est un exemple de rapports bilatéraux. Statistique Canada a tiré de grands avantages du fait de pouvoir utiliser et, dans certains cas, développer des méthodes statistiques mises au point par le Census Bureau, dont celles que décrit Bailar; et, je pense, le Census Bureau n'a pas moins profité des progrès méthodologiques réalisés à Statistique Canada.

En ce qui concerne les relations multilatérales, plusieurs organisations offrent une tribune où les organismes statistiques d'État peuvent échanger leurs vues: l'Organisation des Nations-Unies et ses organismes régionaux et spécialisés, l'Institut international de statistique, particulièrement sa section des statisticiens d'enquêtes et sa section des statistiques officielles, et les associations professionnelles de statisticiens de divers pays. En outre, le Census Bureau et Statistique Canada ont institué des conférences de recherches ou symposiums annuels où peuvent être communiqués les progrès et les résultats d'expériences. En somme, ces rapports bi et multilatéraux remplissent bien leur fonction, qui est de faire en sorte que les contributions d'un organisme à la méthodologie statistique – et beaucoup d'organismes apportent des contributions importantes – sont librement partagées et utilisées par d'autres organismes. Mais quelle a été l'influence de ces développements sur la discipline de la statistique à l'extérieur des organismes d'État? Ici, nous allons nous arrêter aux exemples cités par Bailar pour illustrer son propos, encore qu'il y ait beaucoup d'autres domaines (dont certains sont énumérés à la section 4) auxquels des arguments semblables pourraient s'appliquer. Dans le cas de l'échantillonnage, l'influence sur la statistique comme discipline a été d'une grande portée. L'échantillonnage appliqué à une population finie est aujourd'hui une matière qui fait partie du programme d'enseignement de la statistique dans beaucoup d'universités et à laquelle sont consacrés de nombreux manuels. Les progrès méthodologiques réalisés dans les organismes d'État ont été assimilés et développés par la profession statistique. En fait, quelques-uns pourraient considérer qu'à certains égards ces développements ont été raffinés bien au-delà des besoins pratiques des responsables d'enquêtes. Dans le cas des erreurs non dues à l'échantillonnage, le problème est différent. Les contributions dans ce domaine n'ont pas encore conduit à un ensemble cohérent de théories et de méthodes. Cela ne veut pas dire qu'il n'y ait pas eu de progrès. Au contraire, beaucoup de travaux ont été faits sur cette question. Mais la plupart concernent une enquête en particulier. Ils ont, on peut l'espérer, amélioré de nombreuses enquêtes, servi à documenter beaucoup d'expériences et donné aux praticiens d'utiles enseignements. Mais cette question des erreurs non dues à l'échantillonnage n'a pas encore trouvé sa place dans la statistique comme discipline. En fait, les moyens accrus que nous ont donnés les travaux s'y rapportant trouvent souvent leur application dans des domaines particuliers (sociologie, démographie, etc.) plutôt que dans la statistique proprement dite.

Le problème de la désaisonnalisation est encore un autre cas. Technique à caractère plutôt empirique utilisée par les organismes statistiques, elle est de plus en plus étudiée depuis quelques années par des chercheurs qui tentent de lui donner un fondement solide dans la théorie statistique. Bailar mentionne un certain nombre de questions fondamentales actuellement étudiées à propos d'objectifs et de normes pour la désaisonnalisation. On étudie

COMMENTAIRES

G.J. BRACKSTONE¹

1. Introduction

Cet article montre quelle contribution importante a été celle du U.S. Bureau of the Census, depuis 50 ans, au développement de la méthodologie statistique. Les quatre exemples choisis par Bailar pour illustrer cet apport sont remarquables tant par leur importance intrinsèque que par leur diversité. Ces exemples ne sont pas des variantes d'une percée méthodologique unique, mais des contributions fondamentales dans quatre domaines différents. Peut-être ces exemples illustrent-ils la grande diversité et la difficulté des problèmes méthodologiques que doit vent résoudre les organismes statistiques – diversité et difficulté qui feraient mentir ceux qui pourraient penser que dans le secteur public la statistique n'a rien que de routinier et de banal. Dans la description de ces exemples, l'aperçu sur les circonstances dans lesquelles ont eu lieu ces développements offre un intérêt particulier. Les perfectionnements méthodologiques obtenus ont engendré des solutions dont la généralité dépasse de beaucoup l'envergure des problèmes qui les ont suscités au départ, mais les processus même qui les sous-tendaient méritent d'être examinés de près car ils permettent de déterminer quelles circonstances sont nécessaires pour que de telles percées soient possibles. Je reviendrai sur ce point.

Pendant cette même période de cinquante ans, le U.S. Bureau of the Census a également été très actif dans le domaine de l'automatisation des opérations statistiques. Premier à avoir mis au point des triuses et des tabulatrices de cartes perforées au début du siècle, le Bureau of the Census a aussi été le premier organisme statistique à utiliser un ordinateur, dans les années 50. Dans les années 60, c'est encore lui qui a innové dans l'automatisation de la saisie des données en mettant au point le dispositif FOSDIC permettant de lire une version sur micro-fiche d'un questionnaire porteur de marques. Les innovations du Bureau of the Census sont donc manifestement présentes dans beaucoup d'aspects des tâches d'un organisme gouvernemental de statistique.

2. Diffusion des progrès méthodologiques

Chacune des contributions à la méthodologie décrites par Bailar est née de la nécessité pour un organisme statistique de résoudre un problème pratique réel. La nécessité de recueillir des données supplémentaires à un coût raisonnable et dans un délai acceptable a motivé le développement des méthodes d'échantillonnage probabiliste; la nécessité d'améliorer la qualité des données en tâchant de comprendre, de mesurer et de réduire les erreurs non dues à l'échantillonnage a suscité des travaux dans ce domaine; les progrès au chapitre de la désaisonnalisation semblaient avoir été déterminés par le besoin d'accélérer et de normaliser une méthode manuelle complexe; le problème de la définition d'un processus rationnel et efficace pour assurer la confidentialité des renseignements individuels dans les résultats statistiques a inspiré la recherche dans le domaine de la protection du secret statistique. Les nombreux autres exemples qui auraient pu être cités ont pour caractéristique commune d'avoir eu leur point de départ dans un problème pratique réel.

L'organisme statistique qui met au point des méthodes statistiques conçues pour résoudre des problèmes comme ceux qui viennent d'être énumérés en tire évidemment un avantage immédiat. Mais des contributions de ce genre ont-elles eu un intérêt plus large? Ont-elles fait progresser l'ensemble de connaissances et de méthodes qu'on appelle la statistique? On peut dire que ces développements ont eu des avantages significatifs et globaux pour les organismes statistiques chargés de la production de données sociales et économiques, mais que leur importance pour la statistique comme discipline universitaire, bien qu'elle augmente, n'a pas été aussi grande qu'elle aurait pu l'être.

¹ G.J. Brackstone, statisticien en chef adjoint, Statistique Canada, Ottawa, Ontario.

conseillent vivement d'intégrer les méthodes chronologiques aux méthodes d'estimation afin d'obtenir des résultats plus précis. Il sera intéressant de voir si cette intégration se fera et comment elle se fera.

Les organismes statistiques tentent par la modélisation de produire des données pour petites régions. Alors que les données sont le plus souvent recueillies pour des unités géographiques relativement étendues comme les Etats, on aurait besoin de données pour des unités géographiques plus petites comme les comités. Des conférences ont permis d'évaluer et de comparer diverses méthodes de production de données régionales. Au cours de la dernière décennie, le Census Bureau a pu établir des estimations de la population à l'aide de méthodes empiriques. Les recherches faites sur le sous-dénombrement au Census Bureau ont permis d'examiner plusieurs modèles et ont apporté beaucoup d'éléments pour mieux comprendre le problème.

On est en train d'examiner minutieusement des méthodes spéciales de contrôle et d'imputation et d'élaborer des modèles mathématiques. La tendance à la modélisation s'accroîtra sûrement dans l'avenir.

En conclusion, nous croyons que l'avenir sera propice à la modélisation. Non pas que nous cherchions à dénigrer les méthodes empiriques qui sont en usage actuellement; les statisticiens ont toujours affirmé que théorie et pratique vont de pair. Les méthodes empiriques qui semblent donner des résultats satisfaisants mènent à l'établissement de modèles et à la définition de nouveaux principes théoriques qui sont nuancés par la pratique. Du fait que les organismes d'Etat sont appelés à résoudre de nombreux problèmes statistiques qui pré-sentent le plus grand intérêt, depuis toujours ils font figure de pionnier dans le domaine de la méthodologie statistique.

BIBLIOGRAPHIE

- BAILLAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- BELL, W.R., et HILLMER, S.C. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-317.
- CAUSEY, B.E., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- CITRO, C.F., et COHEN, M.L. (éds.) (1985). *The Bicentennial Census*. Washington, D.C.: National Academy Press.
- DUNCAN, J., et SHELTON, W. (1978). *Revolution in United States Government Statistics*. Washington D.C.: U.S. Government Printing Office.
- HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.
- HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., et BERSHAD, M.A. (1961). Measurement errors in censuses and surveys. *Proceeding of the International Statistical Institute*, 38, 358-374.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W. (1953). *Sample Survey Methods and Theory*, Vols. 1 et 2. New York: John Wiley and Sons.
- OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 191-210.
- U.S. BUREAU OF THE CENSUS (1968). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders*. Series ER 60 n° 7.
- U.S. BUREAU OF THE CENSUS (1979). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1970: Enumerator Variance in the 1970 Census*. PHC(E) n° 13.

On considère qu'il y a divulgation dans un tableau de fréquences lorsque il est possible de conclure que le nombre de répondants visés par une case X quelconque est inférieur à un seuil préétabli. À l'occasion du recensement décennal de 1980, on avait défini des seuils différents pour les ménages et pour les individus.

Les méthodes de protection du secret statistique pour les tableaux de fréquences se divisent en trois catégories: la suppression de cases, la modification de fréquences de case et le remplacement de fréquences de case par des intervalles. La suppression de cases empêche le dévoilement de valeurs numériques et rend impossible toute inférence qui pourrait découler d'une tentative pour établir des relations linéaires entre les fréquences de cases publiées et les fréquences non publiées. La modification de fréquences de case consiste à modifier légèrement à la hausse ou à la baisse la plupart des fréquences de manière que l'on ne puisse tirer de conclusions précises sur les valeurs figurant dans le tableau. La troisième méthode, qui consiste à remplacer les estimations ponctuelles par des intervalles, est peu utile dans le cas de classifications croisées.

La suppression de cases a été la principale méthode utilisée par le Census Bureau au cours des années 1980. Les restrictions appliquées aux sommes de ligne ou de colonne d'un tableau de fréquences produisent une série de contraintes linéaires. Une fois que l'on a supprimé les cases les plus susceptibles de divulguer des renseignements confidentiels, on utilise la programmation mathématique pour vérifier si l'on n'aurait pas d'autres sources de divulgation dans le tableau. Bien que cette méthode ait été utilisée de façon empirique pendant de nombreuses années, Cox et ses collègues du Census Bureau ont défini les fondements mathématiques de la méthode (Caussey, Cox et Ernst, 1985) et ont montré jusqu'à quel point la suppression de cases était une opération complexe.

Les méthodes de modification de fréquences de case, y compris l'arrondissement aléatoire, ont été élaborées et appliquées au Royaume-Uni, en Suède et au Canada. Toutes ces méthodes consistent à ajouter à des fréquences de case ou à soustraire de ces fréquences, avec une probabilité définie, une faible valeur pouvant parfois être zéro.

En ce qui concerne des données comme les chiffres de vente, la valeur nette, les stocks et les données financières d'établissements manufacturiers et d'établissements de vente au détail, le danger, selon le Census Bureau, est que l'on réussisse à découvrir les montants déclarés par un répondant. Si les dirigeants d'une entreprise venaient à déduire des chiffres d'un tableau ceux qui se rapportent à leur entreprise, ils pourraient découvrir par soustraction les chiffres fournis par un concurrent. Pour éviter cela, le Census Bureau a recours à la suppression de cases. À cette fin, il applique la règle dite de (n, k) , selon laquelle X est considérée comme une case prêtant à divulgation si un nombre n de répondants représentent plus de k pour cent de la fréquence totale de la case. Cette règle fait partie d'une série de règles touchant la prédominance des cases et qui sont toutes additives.

La protection du secret statistique est maintenant un sujet de préoccupation dans tous les pays et plus particulièrement au sein des administrations publiques. En effet, celles-ci doivent chercher à résoudre les problèmes très délicats que pose depuis quelque temps la demande de microdonnées.

6. PERSPECTIVES D'AVENIR

L'élaboration de modèles mathématiques joue un rôle important dans les quatre cas que nous venons de traiter. L'échantillonnage repose bien sûr sur des méthodes de randomisation mais la volonté de limiter l'erreur totale dans les enquêtes a favorisé l'élaboration d'un modèle d'erreurs d'enquête énoncé pour la première fois par Hansen, Hurwitz et Berstad (1961). Ce modèle ainsi que les tests ayant servi à estimer les paramètres ont été à l'origine de nombreux décisions concernant les modalités d'exécution des recensements et des enquêtes.

L'usage des modèles de séries chronologiques est très répandu dans le monde; ceux-ci tendent d'ailleurs à remplacer les méthodes empiriques telle la $X-11$. À l'heure actuelle, les spécialistes

de prendre au mot le Federal Reserve Board. Sa proposition allait dans le sens suivant: le Federal Reserve Board choisirait une série quelconque et prendrait tout le temps voulu pour la corriger; de son côté, Shiskin traiterait la même série par ordinateur. Les deux versions ainsi corrigées seraient reproduites graphiquement, sans que l'on précise la façon dont elles ont été corrigées, et soumises à l'attention d'un petit groupe de spécialistes du Federal Reserve Board pour qu'il se prononce sur la qualité des résultats. Les membres de ce groupe ont été unanimes pour dire que la série corrigée par ordinateur était supérieure à l'autre.

Aujourd'hui, les organismes fédéraux désaisonnalisent des milliers de séries chronologiques à chaque année. Pendant de nombreuses années, on a cru que les méthodes fondées sur un modèle étaient difficilement applicables à cause des limites des ordinateurs. Par ailleurs, à chaque année on élaborait de nouveaux facteurs de désaisonnalisation à partir de données chronologiques. Par exemple, un facteur qui devait servir à calculer les données désaisonnalisées pour juillet était élaboré en décembre de l'année précédente. Dans les circonstances, on ne pouvait utiliser de données fondées sur des événements plus récents. Cette restriction était tout à fait logique compte tenu de ce que la préparation des cartes perforées et le traitement de la série à l'ordinateur étaient des opérations qui s'étendaient sur plusieurs jours. Cependant, au cours des dix dernières années cette méthode a fait l'objet de plus en plus de critiques au profit de la méthode de désaisonnalisation courante. Les employés du Census Bureau qui étaient affectés aux séries chronologiques, avec David Findley en tête, étudiaient en profondeur les avantages de cette méthode de désaisonnalisation pour les séries du Census Bureau et prônerent son utilisation au sein du Bureau.

Ces mêmes employés se posèrent aussi des questions très fondamentales en ce qui a trait à la désaisonnalisation. Premièrement, sur quel critère doit-on se fonder pour déterminer si une série doit être désaisonnalisée ou non? Deuxièmement, comment évalue-t-on ces diverses méthodes pour corriger les séries chronologiques, comment évalue-t-on ces diverses méthodes? Dans un article de fond, Bell et Hillmer (1984) s'interrogent sur les nécessités d'une désaisonnalisation lorsque la série en question peut être modélisée convenablement. Ils définissent également des critères permettant d'évaluer les opérations de désaisonnalisation. Il convient de souligner que le Census Bureau n'est pas le seul organisme d'Etat à avoir fait des recherches inédites dans ce domaine. Soit dit en passant, une des grandes réussites de l'équipe du Census Bureau affectée aux séries chronologiques est la série de conférences qu'elle organise régulièrement à l'intention des spécialistes oeuvrant au sein de l'administration publique. Ainsi, des membres du Federal Reserve Board, du Bureau of Labor Statistics, de l'Energy Information Administration et du Bureau of Economic Analysis, pour ne nommer que ceux là, assistent régulièrement à ces conférences pour se tenir au courant des progrès dans le domaine. Estella Dagum, de Statistique Canada, a mené à bien de nombreux projets de recherche, notamment l'élaboration de la méthode X-11 ARMMI.

5. PROTECTION DU SECRET STATISTIQUE

Que l'on soit d'accord ou non avec les lignes directrices du Census Bureau en matière de confidentialité des données, il faut reconnaître que le Bureau a été l'un des premiers à prôner l'utilisation de méthodes visant à protéger la confidentialité des données d'enquête. La protection du secret statistique est une mesure qui vise à garder confidentiels les renseignements fournis par un répondant. La question de la protection du secret statistique a toujours été un problème dans les recensements, mais elle l'est aussi dans les enquêtes, particulièrement les enquêtes de nature longitudinale ou celles où on risquerait d'associer des enregistrements à des données de l'enquête. La principale préoccupation des responsables de recensements démographiques est d'éliminer les risques de divulgation liés à la publication de très petites fréquences. De trop petites fréquences risquent de dévoiler l'identité de répondants ou de petits groupes de répondants. En outre, la présence de zéros dans certaines cases est aussi une source potentielle de divulgation.

enregistrer ses réponses, par un "recensement par la poste", où chaque ménage reçoit un questionnaire par la poste, le remplit et le retourne par la poste. L'expérience a été tentée à l'occasion des recensements de 1960 et de 1970 et on a observé une forte diminution de la variance lorsqu'il y avait auto-dénombrement (U.S. Bureau of the Census 1960, 1970).

Par ailleurs, Hansen et Hurwitz ont encouragé la recherche sur l'erreur de couverture. Le Census Bureau a passé beaucoup de temps à analyser les effets de l'erreur de couverture dans les recensements comme dans les enquêtes. Après le recensement de 1950, le Census Bureau a réussi à mesurer le niveau de sous-dénombrement au niveau national selon l'âge, l'origine raciale et le sexe en se servant d'un modèle qu'avait élaboré Ansley Coale à l'université Princeton. Cette méthode, connue sous le nom d'analyse démographique, a permis de constater que le niveau de sous-dénombrement était beaucoup plus élevé chez les noirs que chez les blancs (Citro et Cohen, 1985). De plus, le Census Bureau a entrepris l'élaboration d'une enquête postcensitaire dans le but de mieux connaître la population non dénombrée. À l'origine, le Bureau cherchait surtout à perfectionner ses méthodes de recensement; depuis quelques années, il met plutôt l'accent sur la répétition d'enquêtes ou de recensements et c'est justement l'approche qu'il entend adopter pour le recensement de 1990. De même, le sous-dénombrement observé dans des enquêtes a poussé les spécialistes à approfondir les méthodes d'estimation par quotient dans l'espoir de réduire l'impact du sous-dénombrement. Ces méthodes sont utilisées dans la plupart des enquêtes-ménages du Census Bureau.

Le U.S. Bureau of the Census est maintenant reconnu pour ses recherches sur les erreurs de mesure. En plus de ses recherches sur les erreurs de réponse et la couverture, il a favorisé la réalisation d'études sur les biais de renouvellement, qui ont un effet sur les estimations tirées d'enquêtes où les répondants sont sollicités plus d'une fois. L'enquête sur la population active, où les répondants demeurent dans l'échantillon quatre mois consécutifs, puis en sont exclus pendant huit mois, et enfin y sont inclus de nouveau pendant quatre mois, a été analysée soigneusement. Bailar (1975) a montré que les estimations du niveau d'emploi et du niveau de chômage étaient en règle générale plus élevées pour les personnes qui en étaient à leur premier mois dans l'échantillon que pour celles qui n'en étaient pas à leur première présence. Ces écarts influent sur les niveaux d'emploi et de chômage, mais n'ont probablement aucun effet sur les estimations de la variation d'un mois à l'autre.

Ceci n'est qu'un aperçu des recherches qui ont été entreprises au Census Bureau sur les erreurs de mesure. À l'heure actuelle, tous les organismes statistiques s'intéressent à la question.

4. DÉSAISONNALISATION

On a commencé à parler de désaisonnalisation au sein de l'administration publique lorsque Julius Shiskin était membre du Census Bureau. Il était alors chargé d'informer les opérations de désaisonnalisation. Aujourd'hui, la méthode X-11 est utilisée partout. Shiskin raconte que dans les années 1950, le Council of Economic Advisors pressait les organismes fédéraux de produire des séries chronologiques désaisonnalisées. En 1953, le Census Bureau acquiert le premier ordinateur spécialisé dans le traitement des données, le UNIVAC I, et Eli Marks, qui se rend au travail dans la même voiture que Shiskin, l'entretenant souvent de la difficulté qu'il éprouve à programmer cet ordinateur. Shiskin se demande s'il ne serait pas possible de se servir de l'ordinateur pour exécuter la désaisonnalisation des séries; après avoir consulté un technicien en informatique, il découvre qu'une série de 10 ans pourrait être désaisonnalisée en une minute. Aujourd'hui, une opération de ce genre prend évidemment beaucoup moins de temps.

La désaisonnalisation est en quelque sorte une opération qui exige de la maîtrise étant donné le très grand nombre d'options qu'offre le programme X-11 à l'analyste. Cependant, lorsque Shiskin a entrepris d'automatiser les opérations de désaisonnalisation, on se demandait si une machine pouvait vraiment effectuer le travail d'un technicien expérimenté. Shiskin décida donc

qu'on donne à l'enquête sur la population active) a servi de modèle dans le monde entier pour l'élaboration d'enquêtes sur la population active.

La Statistical Research Division s'est aussi intéressée aux problèmes d'échantillonnage que posaient les enquêtes menées auprès d'établissements commerciaux. L'idée la plus courante était que l'échantillonnage était une opération appropriée pour des populations relativement homogènes comme les populations de personnes mais convenait peu pour des populations fortement asymétriques comme les populations d'entreprises. Tenant compte de l'asymétrie de la population étudiée, le groupe de recherche a stratifié les magasins de détail selon leur taille. Les plus gros magasins étaient nécessairement inclus dans l'échantillon et ceux de moindre importance étaient échantillonnés avec une probabilité proportionnelle à la taille.

On a aussi observé qu'il y avait de nombreuses créations et disparitions d'entreprises. Comme on croyait qu'un échantillon statique ne pourrait pas refléter cette rotation, on a créé un échantillon aréolaire qui allait produire des estimations pour les nouveaux magasins. Bien que l'enquête mensuelle sur le commerce de détail ait subi de nombreux autres perfectionnements, sa structure fondamentale est demeurée la même. L'estimation composite est aussi utilisée dans l'enquête sur le commerce de détail afin d'obtenir des estimations plus précises.

Nous pourrions citer beaucoup d'autres exemples de perfectionnement des méthodes de sondage. L'ouvrage en deux volumes de Hansen, Hurwitz et Madow intitulé *Simple Survey Methods and Theory* contient de nombreux exposés accompagnés de la théorie et des applications correspondantes. Bien que les exemples qui y sont présentés soient largement périmés, nous ne connaissons pas d'ouvrage qui contienne autant d'applications relatives aux sondages. Le seul inconvénient, c'est que ces applications n'aient été jamais mises à jour.

3. ERREUR NON DUE À L'ÉCHANTILLONNAGE

Dans cet effort soutenu pour améliorer les méthodes de recensement et d'enquête, il ne fallait pas s'arrêter uniquement aux erreurs d'échantillonnage. Il fallait en effet tenter de limiter les erreurs provenant d'autres sources, comme l'interview, le traitement, le questionnaire, etc. Hansen et Hurwitz ont commencé à s'intéresser à la question avant le recensement de 1950; ils ont intégré à ce recensement de nombreuses études expérimentales qui visaient à estimer l'effet des erreurs de mesure. L'erreur totale dans les enquêtes devint donc un sujet de recherche important au Census Bureau. L'évaluation et le contrôle des erreurs non dues à l'échantillonnage figuraient régulièrement au programme de recherche du Census Bureau.

L'assertion voulant que les erreurs de mesure pourraient avoir une influence beaucoup plus marquée sur les données que les erreurs d'échantillonnage, surtout aux niveaux d'aggrégation supérieurs, a donné un nouvel élan à la recherche sur les erreurs non dues à l'échantillonnage. Hansen, Hurwitz et Bershad (1961) ont élaboré un modèle intégré pour les recensements et les enquêtes, qui tenait compte à la fois de l'erreur d'échantillonnage, de l'erreur de réponse et du biais. L'erreur de réponse était composée de ce qu'on appelle aujourd'hui la variance de réponse simple et la variance de réponse corrélée. La première représente la variabilité fondamentale observée d'une fois à l'autre à cause des différences entre les réponses, les répondants, les interviewers, etc. La variance de réponse simple représente aussi la variabilité observée d'une fois à l'autre dans le codage. La variance de réponse corrélée représente la variance engendrée par un facteur qui tend à orienter les réponses. Le facteur auquel on s'intéresse le plus souvent est l'interviewer. Parce qu'il peut avoir une certaine idée de la réponse que fournira la personne interviewée ou parce qu'il a acquis de l'expérience dans l'interview des ménages, l'interviewer peut orienter les réponses vers certaines catégories. Pour une région donnée, on observe une forte variabilité selon les interviewers des taux de non-réponse et des réponses concernant le niveau de scolarité et beaucoup d'autres caractéristiques.

Le modèle de Hansen, Hurwitz et Bershad a été testé pour la première fois lors du recensement de 1950 et il a eu une influence déterminante dans la décision de remplacer le "recensement sur place", où un interviewer visitait chaque ménage pour lui poser des questions et

gros établissements. Toutefois, les méthodes d'échantillonnage n'avaient à peu près pas de fondements théoriques. En 1937, le Congrès a autorisé un recensement facultatif des personnes en chômage et des personnes employées à temps partiel dans tout le pays. Un questionnaire devait être expédié par la poste à chaque ménage. Comme on craignait que ce recensement facultatif renferme une erreur systématique quelconque, on a décidé de procéder à un recensement de contrôle dans un certain nombre de secteurs. Ce recensement consistait à former un échantillon probabiliste de routes postales et à interviewer tous les ménages dont le logement se trouvait sur ces routes. Les facteurs étaient chargés de faire les interviews et de classer les questionnaires qui étaient retournés par la poste. Ils produisaient ensuite des chiffres pour chaque route postale, y compris les routes échantillonnées. On disposait ainsi d'une variable indépendante pour l'estimation; c'était le début de l'estimation par quotient. Les résultats du recensement de contrôle ont confirmé l'utilité de l'échantillonnage. Cependant, l'expérience a été remarquable à plusieurs égards:

- évaluation précise des effets de la non-réponse pour un recensement facultatif;
- utilisation de l'estimation par quotient;
- production rapide des résultats.

Au cours d'une interview reproduite dans la revue *Statistical Science* (Olkinn), Hansen raconte que le recensement facultatif s'est fait durant la semaine du 20 novembre 1937, que les interviews ont eu lieu dans la semaine du 4 décembre et que les résultats préliminaires ont été connus le 31 décembre. Il est difficile de croire que le Census Bureau puisse agir encore plus rapidement aujourd'hui. Selon Hansen, les résultats convainquants du recensement de contrôle de 1937 auront eu le mérite de faire avancer la cause de l'échantillonnage probabiliste au sein du Census Bureau. Auparavant, on croyait au Census Bureau qu'il fallait s'en tenir à des recensements et que les sondages n'étaient pas un exercice sérieux. Grâce au succès de l'expérience de 1937, on pouvait envisager de recourir à l'échantillonnage à l'occasion du recensement de 1940; celui-ci fut d'ailleurs le premier recensement où certaines questions étaient adressées uniquement à un sous-ensemble de la population. Malheureusement, quelques membres du Census Bureau ont repris depuis quelques mois l'idée de faire une vérification complète des logements inoccupés sous prétexte que l'un recensement comporte moins d'erreurs qu'une enquête. Espérons qu'il ne s'agit là que d'un moment d'aberration causé par un différend quelconque.

La théorie des sondages a évolué au même rythme que l'enquête sur la population active. La Works Progress Administration (WPA) avait la responsabilité d'une enquête visant à mesurer le niveau de chômage. Lorsque cet organisme a cessé d'exister en 1942, le Census Bureau s'est vu confier la responsabilité de l'enquête. Il a alors fait une évaluation des méthodes de sondage et leur a apporté de nombreuses améliorations. Cet exercice de révision a contribué largement à l'avancement de la théorie des sondages. Mentionnons au passage quelques-uns des principes qui ont été définis lors de cet exercice de révision: unités primaires d'échantillonnage plus grandes, échantillonnage avec probabilité proportionnelle à la taille et sous-stratification de régions ou de secteurs. Hansen et Hurwitz ont analysé ces principes dans un article publié en 1943 dans la revue *Annals of Mathematical Statistics*. Lorsqu'on relit "On The Theory Of Sampling From Finite Populations", on découvre toujours quelque chose de nouveau. Ce serait le premier article que des employés d'un organisme fédéral auraient publié sur l'échantillonnage appliqué à des populations finies. Bien que les notions aient déjà été traitées par d'autres auteurs, ce que Hansen et Hurwitz ont été les seuls à faire une analyse des résultats à l'aide d'une série de comparaisons qui faisaient ressortir les avantages des méthodes proposées. Par la suite, on a continué d'améliorer l'enquête sur la population active. Ainsi est arrivée l'estimation composite, qui applique le principe du renouvellement de l'échantillon afin d'obtenir de meilleures estimations. Nul doute que la Current Population Survey (c'est le nom

Rôle de l'administration fédérale dans le développement des méthodes statistiques aux États-Unis

BARBARA A. BAILLAR¹

RÉSUMÉ

Dans cet article, nous montrons brièvement comment les statisticiens du U.S. Bureau of the Census ont pu, par leurs recherches, contribuer à l'avancement de la théorie et de la pratique des recensements et des sondages. Nous essayons aussi de voir ce que nous réserve l'avenir à ce chapitre.

MOTS CLÉS: Échantillonnage; erreur non due à l'échantillonnage; estimation; confidentialité; désaisonnalisation.

1. INTRODUCTION

Aux États-Unis, l'administration fédérale a été un chef de file dans l'élaboration de méthodes statistiques pour les recensements et les sondages. Nous nous limiterons ici aux réalisations du U.S. Bureau of the Census et concentrerons notre attention sur quatre grands sujets de recherche - l'élaboration de méthodes d'échantillonnage, l'erreur non due à l'échantillonnage, la désaisonnalisation et l'élaboration de méthodes visant à protéger le caractère confidentiel des données fournies par les répondants (communément appelées méthodes de protection du secret statistique). Enfin, nous tenterons de voir comment pourrait évoluer la recherche dans les années à venir.

2. ÉCHANTILLONNAGE

L'histoire des méthodes de sondage au sein de l'administration fédérale des E.-U. est sur tout celle d'un groupe de personnes remarquables qui ont oeuvré au U.S. Bureau of the Census sous la direction de Morris Hansen et de William Hurwitz. Lorsqu'on apprend que le U.S. Bureau of the Census était acquis à l'échantillonnage probabiliste dès le début des années 1940, on se demande comment une institution aussi conformiste a pu faire pour adopter si tôt une telle position? En règle générale, les organismes hésitent très longuement avant d'adopter de nouvelles méthodes et probablement que le Census Bureau est beaucoup moins empressé aujourd'hui à adopter de nouvelles méthodes et à en promouvoir l'utilisation. Hansen donne trois raisons pour expliquer que les divisions spécialisées du Census Bureau aient accepté assez rapidement l'idée de l'échantillonnage (Causey, Cox et Lawrence 1985) : attitude favorable de la part de la direction du Census Bureau (Bailar 1975), établissement de liens de coopération avec les divisions spécialisées (Bell et Hillmer 1984) et formation d'un groupe d'experts en sondages (appelés ultérieurement méthodologistes) au sein des divisions spécialisées, ce groupe étant chargé de conseiller la Statistical Research Division (SRD) sur les questions d'ordre technique. Hansen omet de mentionner un autre facteur important, soit la vigueur et le caractère du duo dynamique qu'il forme avec Hurwitz et de ses partisans.

En 1936, le Census Bureau a commencé à s'intéresser à l'échantillonnage et aux applications possibles. À cette époque, on pratiquait déjà l'échantillonnage mais il ne s'agissait pas d'échantillonnage probabiliste. Il y avait l'échantillonnage par choix raisonné et l'échantillonnage de

¹ Barbara A. Bailar, American Statistical Association, 1429 rue Duke, Alexandria, VA 22314-3402.

Il aurait donc été intéressant que les auteurs fassent des comparaisons internationales pour mettre en évidence l'histoire de la recherche dans le domaine des enquêtes dans diverses sociétés.

4. Large acception du modèle socio-psychologique de l'interview dans les milieux universitaires à partir de 1960

Dans ce contexte, l'interview est habituellement décrite comme étant "une conversation dirigée", et l'attention du chercheur est axée sur le rôle des deux protagonistes lorsqu'il étudie les erreurs produites lors du sondage.

5. Ubiquité des enquêtes

Les sondages font actuellement partie des activités courantes de la plupart des grosses sociétés (avant l'abolition du monopole d'AT&T aux Etats-Unis, cette société procédait chaque année à plus de 7 millions d'interviews pour mesurer le degré de satisfaction des consommateurs). Les sondages sont considérés comme étant des sources d'information irremplaçables sur les clients, les fournisseurs et la société en général.

6. Non-réponse et répugnance croissante de la population à faire l'objet de sondages

Dans la plupart des pays occidentaux, ce phénomène révèle une grande importance pour les chercheurs qui s'intéressent aux enquêtes. Le fait que les déductions statistiques sont applicables à de grandes populations étant l'une des principales vertus des sondages par rapport à d'autres modes de collecte des données, cette question frappe au coeur du sujet. Là encore, si l'article avait comparé la situation dans divers pays, il aurait jeté la lumière sur ces questions. Nous pouvons appliquer l'idée de superpopulation à tout récit historique; autrement dit, toute série d'événements (qui, plus tard, forment l'"histoire") n'est qu'une réalisation d'un ensemble infini de séries possibles qui déterminent l'univers des réalités possibles. Cette réflexion concerne l'ensemble de questions qui demeurent sans réponse:

1. Pourquoi, après presque un siècle, la recherche dans le domaine des enquêtes n'a-t-elle pas donné naissance à une profession (avec des normes et des critères de formation précis)?
2. Pourquoi y a-t-il si peu de structures dans l'enseignement régulier permettant aux chercheurs d'acquérir les connaissances de base? Pourquoi les universités n'ont-elles pas prévu de départements de recherche dans le domaine des enquêtes? Pourquoi y a-t-il des départements de communications, de recherche opérationnelle et d'architecture navale, mais aucun département de recherche dans le domaine des enquêtes (où l'on enseignerait l'échantillonnage, la conception des questionnaires, l'analyse des données)?
3. La sensibilisation du public à l'égard des sondages et des statistiques (comme le programme de l'ASA/NSF qui vise à familiariser le public avec les calculs) aurait-elle eu une incidence sur l'acceptation des sondages?

Nous sommes redevables à l'équipe Fienberg-Tanur de l'examen de notre passé. Ils ont contribué à faire la chronique de la naissance et des cinquante premières années de ce qui est maintenant une composante importante de la plupart des sociétés du monde. J'espère sincèrement qu'en 2040, à l'occasion de cet autre anniversaire, il sera nécessaire de demander à Fienberg et Tanur de mettre à jour leur article. J'espère également qu'ils seront à même de signaler des innovations survenues au cours de ces cinquante années qui auront amélioré les méthodes de sondage.

début de l'histoire de la méthode (l'équipe de Likert n'a pu obtenir des vignettes de stationnement à l'université parce qu'elle n'en faisait pas vraiment partie). Même à présent, dans de nombreuses cités universitaires, cette recherche est souvent considérée comme un refuge pour les techniciens (situés plusieurs degrés au-dessous des laborantins en chimie). En revanche, certains organismes gouvernementaux et des sociétés d'études de marché se consacrent entièrement à l'élaboration des plans de sondage, ainsi qu'à la collecte et à l'analyse des données. On y trouve partout des décideurs qui surveillent continuellement la structure des coûts et des erreurs des sondages sans s'engager dans l'éternel débat sur la valeur relative de l'entreprise. L'article s'achève sur l'étude de trois progrès réalisés depuis 1960 qui sont importants si l'on veut comprendre les sondages. À ce moment, l'article cesse d'être axé sur les institutions et s'articule plutôt autour des innovations. Trois d'entre elles sont mises en lumière: a) l'utilisation du téléphone comme moyen de collecte des données et les progrès récents en matière d'interview téléphonique assistée par ordinateur (ITAO); b) l'utilisation des enquêtes longitudinales pour étudier les variations de valeurs individuelles dans le temps; et c) l'application des concepts de la psychologie cognitive aux méthodes d'enquête.

Les auteurs prennent note de l'évolution du mode de collecte des données, de l'interview sur place à l'interview téléphonique et à la mise au point de l'ITAO, mais il ne mentionnent pas que, aux États-Unis, c'est en grande partie un phénomène propre aux universités et au secteur public (le secteur privé avait adopté la méthode depuis des années). En fait, c'est un exemple qui illustre l'utilisation de méthodes distinctes par les trois secteurs. Comme eux, je pense qu'on reconnaît de plus en plus les mérites des enquêtes longitudinales et je remarque qu'on l'a fait de plus en plus partout dans le monde au cours des années quatre-vingts. C'est grâce à l'équipe Fienberg-Tanur que les États-Unis ont entrepris d'appliquer les concepts de la psychologie cognitive aux sondages, et nous devons les en remercier. L'article ne nous permet pas de décider si les auteurs estiment que l'ITAO, les enquêtes longitudinales et les efforts en vue de faire intervenir la psychologie cognitive dans la recherche sur les techniques d'enquête sont les trois progrès *les plus* importants dans le domaine; toutefois, il est clair qu'ils en omettent plusieurs autres. Nous pouvons tous choisir les trois progrès les plus importants à notre avis depuis 1960; en voici des exemples:

1. **Elaboration de logiciels statistiques généralisés**

Cette innovation a beaucoup accru le nombre des chercheurs qui pouvaient poser des questions et y répondre directement en utilisant les données d'enquête. Au moment où j'écris, en statistique et dans les sciences sociales, il est courant pour les étudiants du premier cycle d'effectuer des analyses de ces données, ce dont ils auraient été incapables voici 25 ans, en raison de leur complexité.

2. **Existence de fichiers de données d'enquête**

Le stockage des données d'enquête sur support informatique a encore contribué à la démocratisation de l'analyse des résultats des enquêtes. Avec son avènement, la répétition et l'extension de l'analyse, composantes clés de la structure du progrès scientifique, sont devenues banales. Malheureusement, il y a eu aussi des effets nuisibles. Les analystes des données d'enquête pouvaient faire leur travail en ignorant complètement le plan de sondage, la formation de l'intervieweur et les lignes de conduite relatives à la supervision, les taux de non-réponse et une foule d'autres aspects de l'élaboration d'une enquête, connus des enquêteurs.

3. **Croissance du nombre d'entreprises commerciales et d'organismes sans but lucratif effectuant des enquêtes pour le compte des administrations publiques**

C'est une particularité des États-Unis qu'on y recourt à des entreprises commerciales et à des universités pour effectuer des enquêtes pour le compte d'organismes gouvernementaux. Il en est de même dans de nombreux pays occidentaux, mais à une beaucoup plus petite échelle.

à la Columbia University qui obtient un succès partiel à cet égard est instructif. De même, Likert et autres, qui ont quitté un organisme public (le Département de l'agriculture des E.-U.) pour se joindre à une université afin d'étendre la méthode à de nouveaux domaines, nous montrent que cette histoire est avant tout celle d'un certain nombre de groupes de personnes et des organisations qui les ont rendus efficaces.

Toutefois, l'accent mis sur les organismes ou les institutions pourrait laisser penser, à tort, que ceux-ci ont donné une impulsion aux innovations. Or, rien dans l'article ne change mon opinion, à savoir qu'à son origine, le domaine des sondages a attiré des penseurs créatifs qui avaient une vision globale des choses. Nombre d'entre eux étaient intelligents et charismatiques; ils ont ouvert la voie avec leurs idées et ont inspiré des disciples qui se sont attachés à délimiter le nouveau domaine. Les institutions ont permis le déroulement de ces travaux, mais elles n'ont pas généré les progrès; elles se sont bornées à accueillir les éléments les plus éminents et les plus brillants.

J'aurais voulu que l'article, dans l'optique choisie, fasse une plus large place à deux points connexes:

1) Des tâches différentes ont été plus facilement réalisées dans différents domaines. Par exemple, de par leur nature, les organismes gouvernementaux étaient limités à l'étude des questions concernant l'aide sociale, les sociétés d'études de marché, à l'étude des questions d'actualité ou de rentabilité, et les universités, à l'étude des questions sociales plus fondamentales, présentant un intérêt à long terme. Ceux qui ont participé aux premiers travaux ont adapté leur programme aux buts de l'organisation.

2) Les témoignages concernant l'époque où la recherche en était à ses débuts dénotent l'enthousiasme des pionniers. J'ai trouvé que les auteurs de l'article n'ont pas suffisamment montré combien les jeunes chercheurs avaient le sentiment de mener ensemble une mission évangélique: répandre "l'évangile" de l'échantillonnage probabiliste, inventer de nouvelles méthodes d'interview, parce qu'il fallait partir à zéro. En mettant l'accent sur les institutions, les auteurs ont oublié le drame humain qui s'est joué à ce moment-là.

Fienberg et Tanur font également remarquer que "la démarcation entre les institutions n'est pas absolue". Autrement dit, les chercheurs vont et viennent entre les institutions, faisant des apports à chacune d'entre elles au fur et à mesure de leurs déplacements. Les auteurs en veulent pour preuve le cas de Lazarsfeld qui s'est penché sur les aspects fondamentaux de l'établissement des plans de sondage tout en effectuant dans une université des travaux de recherche sur les cotes d'écoute de la radio, et celui de Likert qui est passé du domaine des assurances au Département de l'agriculture et ensuite à l'université du Michigan. Ces déplacements semblent être l'exception plutôt que la règle. Je n'ai pas entrepris les recherches nécessaires sur le cheminement de carrière des personnes en cause pour en faire la preuve, mais j'ai l'impression que les barrières entre ces secteurs ont toujours été et demeurent élevées et qu'on ne peut les franchir sans dommage. De plus, le passage du milieu universitaire aux administrations publiques puis aux sociétés d'études de marché est généralement unidirectionnel. On circule rarement du secteur privé ou public vers le milieu universitaire (à cause des exigences actuelles en matière de publication). Les échanges réciproques entre les administrations publiques et le secteur privé sont plus importants.

Cette insularité mène à l'élaboration de techniques réservées aux différents secteurs et non interchangeables (méthodes de contrôle et d'imputation, méthodes visant à réduire le taux de non-réponse). Dans une certaine mesure, les trois secteurs ont élaboré leur propre langage pour décrire leur travail (par ex. tracé en "tige et feuilles", "totalisation" contre "tableaux de contingence").

De plus, l'idée de démarcation ne tient pas compte des grandes différences qui existent quant à l'importance des sondages dans l'optique des trois secteurs. Dans aucune université au monde la recherche dans le domaine des enquêtes par sondage est-elle cruciale. Elle ne l'était pas au

COMMENTAIRES

ROBERT M. GROVES¹

Le fait qu'on écrive l'histoire de l'élaboration et de l'utilisation des méthodes de sondage indique que le domaine a atteint une certaine maturité. Actuellement, nous célébrons le cinquantième anniversaire de plusieurs innovations importantes en matière de sondage: la parution des documents révolutionnaires de Neyman sur la stratification, l'instauration de la U.S. Current Population Survey et l'usage plus répandu du sondage électoral. Face à ces éléments nouveaux, il est naturel de faire le bilan des années écoulées pour chercher à relier entre eux les événements marquants dans le domaine. C'est à quoi se sont attachés les professeurs Fienberg et Tanur dans leur article.

Je vais passer en revue les principales parties de cet article en présentant mes commentaires au fur et à mesure; j'appellerai ensuite l'attention du lecteur sur des erreurs de non-observation, des éléments mis en valeur à tort et autres petites critiques.

Fienberg et Tanur ont deux façons d'expliquer l'objet de leur article: en disant d'abord que "pour bien comprendre l'évolution des méthodes d'enquête au sens technique, il est nécessaire de suivre l'évolution des établissements chargés de la réalisation des enquêtes" (p. 33), et en faisant remarquer ailleurs que "les progrès de la théorie statistique ne sont pas la seule chose qui ait déterminé l'évolution de la théorie des sondages." (p. 45). En conséquence, ils mettent en évidence:

1. le rôle des institutions dirigeantes qui perçoivent la nécessité d'avoir des informations sur le bien-être de la population ou sur ses réactions aux mesures fiscales;
2. plus tard, le rôle des spécialistes des sciences sociales dans le milieu universitaire qui définissent les questions centrales du point de vue des sondages relativement aux statistiques et aux mesures;
3. le rôle de l'utilisation que les mass media font des sondages à l'occasion des élections et du compte rendu des actualités; et
4. encore plus tard, l'utilisation des enquêtes par les entités commerciales dans l'économie de marché.

Il s documentent le règlement des controverses au sein des administrations publiques quant à l'utilisation du sondage probabiliste.

Au cours de notre lecture, nous apprenons des faits intéressants: par exemple, il n'y avait pas d'organisation permanente telle le Census Bureau lors de douze recensements aux États-Unis (sur une période de 120 années); le Département de l'agriculture a commencé à recueillir des données parce qu'on avait besoin d'informations sur les approvisionnements alimentaires pendant la Guerre civile; la création de programmes gouvernementaux dans le cadre du New-Deal a donné une nouvelle impulsion aux sondages. L'idée qui semble revenir est que les gouvernements qui donnent une grande importance aux services visant le bien-être de la population exigent plus d'informations sur leur société que ceux qui poursuivent d'autres objectifs. De plus, nous voyons que les gouvernements le plus à l'écoute de l'opinion publique réclament davantage de mesures de cette opinion (je me souviens à ce propos de l'analyse que Gallup établit au début de sa carrière entre le sondage et le vote).

La place importante accordée dans l'article au rôle que les établissements chargés de la réalisation des enquêtes ont joué dans l'évolution du domaine n'est justifiée que pour certaines parties de l'examen de la question. Par exemple, ce point de vue est convaincant lorsqu'on nous décrit les efforts méritoires de Lazarsfeld pour amener à collaborer les sociétés d'études de marché et les instituts de recherche universitaires. Le rôle du Bureau of Applied Social Research

¹ Robert M. Groves, The University of Michigan et U.S. Bureau of the Census.

- PAYNE, S.L. (1951). *The Art of Asking Questions*. Princeton: Princeton University Press.
- PORTER, T.M. (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton: Princeton University Press.
- RAO, J.N.K., et BELLHOUSE, D.R. (1990). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *Techniques d'enquête*, ce numéro.
- RICE, S. (1928). *Quantitative Methods in Politics*. New York: Knopf.
- RUGG, D., et CANTRIL, H. (1944) (1947). The wording of questions. Dans *Gauging Public Opinion*, (éd. H. Cantril). Princeton: Princeton University Press, 23-50.
- SCHWARZ, N. (1987). Cognitive aspects of labor surveys in a multinational context. Document préparé pour le Groupe de travail sur les statistiques de l'emploi, OCDE, Paris, avril 1987.
- SCHUMAN, H., et PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic.
- SMITH, T.W. (1975). Social change and the General Social Survey: An annotated bibliography. *Social Indicators Research*, 2, 9-38.
- STASNY, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating gross labor force flows. *Journal of Business and Economic Statistics*, 6, 207-219.
- STIGLER, S.M. (1986). *The History of Statistics. The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.
- SUDMAN, S. (1967). *Reducing the Costs of Surveys*. Chicago: Aldine.
- SUDMAN, S., et Bradburn, N.M. (1974). *Response Errors in Surveys: A Review and Synthesis*. Chicago: Aldine.
- TAUBER, C. (1978). Census. Dans *International Encyclopedia of Statistics*, (eds. W.H. Kruskal et J.M. Tanur) New York: Macmillan and the Free Press, 42-46.
- THORNBERY, O.T. Jr., et MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. Dans *Telephone Survey Methods*, (eds. R.M. Groves et coll.). New York: Wiley, 25-49.
- TREWIN, D., et LEE, G. (1988). International comparisons of telephone coverage. Dans *Telephone Survey Methods*, (eds. R.M. Groves et coll.). New York: Wiley, 9-24.
- VAN KLEECK, M. (1930). The Federal Unemployment Census of 1930. *Proceedings of the American Statistical Association*, 189-200.
- WHITE, A.A., et BERK, M.L. (1987). Recall strategies in personal interviewing: moving results from the laboratory to the field. *Proceedings of the Social Statistics Section, American Statistical Association*, 66-71.
- WILCOX, W.F. (1930). Census. Dans *Encyclopaedia of Social Sciences*, (eds. E.R.A. Seligman et A. Johnson). New York: Macmillan, 295-300.

- HANSEN, M.H., et HURWITZ, W.N. (1942). Relative efficiencies of various sampling units in population inquiries. *Journal of the American Statistical Association*, 37, 89-94.
- HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: Wiley.
- HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., et MAUDLIN, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- HIPPLER, H.-J., SCHWARZ, N., et SUDMAN, S., eds. (1987). *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.
- JABINE, T.B., STRAF, M., TANUR, J.M., et TOURANGEAU, R., eds. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington: National Academy Press.
- JENSEN, A. (1926). Report on the representative method in statistics. *Bulletin de l'Institut International de Statistique*, 22, 359-380.
- KALTON, G., KASPRZYK, D., et DUNCAN, G.J., eds. (1989). *Panel Surveys*. New York: Wiley.
- KIAER, A.N. (1895-1896). Observation et expériences concernant des dénominations représentatives. *Bulletin de l'Institut International de Statistique*, 9, Liv. 2, 176-183.
- KRUSKAL, W.H., et MOSTELER, F. (1980). Representative sampling, IV: the history of the concept in statistics, 1895-1939. *Revue Internationale de Statistique*, 48, 169-195.
- LAZARSFELD, P.F., BERELSON, B., et GAUDET, H. (1944). *The People's Choice: How the Voter Makes up his Mind in a Presidential Campaign*. New York: Columbia University Press.
- LECUYER, B., et OBERSCHALL, A. (1978). Social research, the early history of. Dans *International Encyclopedia of Statistics*, (Eds. W.H. Kruskal et J.M. Tanur). New York: Macmillan and the Free Press, 1013-1031.
- LEVENSTEIN, A. (1912). *Die Arbeitsfrage mit besonderer Berücksichtigung der sozialpsychologischen Seite des modernen Grossbetriebes und der psychophysischen Einwirkungen an die Arbeiter*. (En Allemand) Munich: Reinhardt.
- MADANSKY, A. (1986). On biblical censuses. *Journal of Official Statistics*, 2, 561-569.
- MASSÉY, J.T. (1988). An overview of telephone coverage. Dans *Telephone Survey Methods*, (eds. R.M. Groves et coll.) New York: Wiley, 3-8.
- MOSTELER, F. (1978). Errors: I. Nonsampling errors. Dans *International Encyclopedia of Statistics*, (eds. W.H. Kruskal et J.M. Tanur). New York: Macmillan and the Free Press, 208-229.
- MOSTELER, F., HYMAN, H., MCCARTHY, P.J., MARKS, E.S., et TRUMAN, D.B. (1949). *The Pre-election Polls of 1948*. Bulletin 60. New York: Social Science Research Council.
- NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- NEYMAN, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. Washington DC: Graduate School, Département de l'agriculture des E.-U.
- NICHOLS, W.L. II (1988). Computer-assisted telephone interviewing: A general introduction. Dans *Telephone Survey Methods*, (eds. R.M. Groves et coll.) New York: Wiley, 337-385.
- NICHOLS, W.L. II, et GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.
- NORWOOD, J.L., et EARLY, J.F. (1984). A century of methodological progress at the U.S. Bureau of Labor Statistics. *Journal of the American Statistical Association*, 79, 748-761.
- OLKIN, I. (1987). A conversation with Morris Hansen. *Statistical Science*, 2, 162-179.

- BOURGUET, M.-N. (1988). Décrire, Compter, Calculer: The debate over statistics during the Napoleonic Period. Dans *The Probabilistic Revolution, Volume I, Ideas in History* (éds. L. Kruger, L.J. Daston et M. Heidelberger). Cambridge: MIT Press 303-316.
- BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique*, 22, 6-62.
- BRADBURN, N.M., SUDMAN, S., et ASSOCIÉS (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- BROOKS, C.A., et BAILLAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Working Paper 3, Office of Federal Statistical Policy and Standards. Washington: Département du commerce des E.-U.
- CANTRIL, H. (1944)(1947). *Gauging Public Opinion*. Princeton: Princeton University Press.
- COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- CONVERSE, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.
- DEMING, W.E. (1944). On errors in surveys. *American Sociological Review*, 19, 359-369.
- DUBOIS, W.E.B. (1899) (1973). *The Philadelphia Negro: A Social Study; Together With a Special Report on Domestic Service by Isabel Eaton*. Millwood, N.Y.: Kraus Reprint.
- DUNCAN, J.W., et SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. Département du commerce des E.-U. Washington, U.S. Government Printing Office.
- FATHI, D., SCHOOER, J., et LOFTUS, E. (1984). Moving survey problems into the cognitive survey laboratory. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 19-21.
- FIENBERG, S.E. (1978). Victimization and the National Crime Survey: Problems of design and analysis. Dans *Survey Sampling and Measurement*, (éd. K. Namboodiri). New York: Academic. 89-106.
- FIENBERG, S.E., et TANUR, J.M. (1983). Large scale social surveys: perspectives, problems, and prospects. *Behavioral Science*, 28, 135-153.
- FIENBERG, S.E., et TANUR, J.M. (1986). The design and analysis of longitudinal surveys: Controversies and issues of cost and continuity. Dans *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits* (éds. R.W. Pearson et R.F. Boruch). New York: Springer-Verlag.
- FIENBERG, S.E., et TANUR, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. *Revue Internationale de Statistique*, 55, 75-96.
- FIENBERG, S.E., et TANUR, J.M. (1988). From the inside out and the outside in: combining experimental and sampling structures. *La Revue Canadienne de Statistique*, 16, 135-151.
- FIENBERG, S.E., et TANUR, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, 1017-1022.
- GINI, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population Italienne (1^{er} décembre 1921). *Bulletin de l'Institut International de Statistique*, 23, 198-215.
- GINI, C., et GALVANI, L. (1929). Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione (1^o dicembre, 1921) (En Italien). *Annali di Statistica*, Série 6, 4, 1-107.
- GRAUNT, J. (1662)(1939). *Natural and Political Observations Made Upon the Bills of Mortality* (éd. Willcox, Walter F. avec introduction). Baltimore: Johns Hopkins Press.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLS, W.L. II., et WAKSBERG, J., éds. (1988). *Telephone Survey Methods*. New York: Wiley.
- GROVES, R.M. (1989). *Surveys Errors and Survey Costs*. New York: Wiley.
- GROVES, R.M., et KAHN, R.L. (1979). *Surveys by Telephone*. New York: Academic Press.

découvertes et discute de questions qui intéressent la profession. Le programme de la National Science Foundation sur les méthodes d'évaluation et l'amélioration des données (Measurement Methods and Data Improvement - MMDI), qui est sous la direction de Murray Aborn, a entre autres pour mission de favoriser la collaboration entre l'administration publique et le monde universitaire. À cette fin, le programme prévoit, par exemple, le versement de subventions de recherche à des universitaires pour l'utilisation et l'amélioration des bases de données de l'administration publique (le colloque de 1983 sur les aspects cognitifs des méthodes d'enquête était parrainé par MMDI) ainsi que le financement d'un programme de bourses offert par l'ASA. En vertu de ce programme, des universitaires vont travailler dans les organismes statistiques fédéraux pendant un semestre ou une année complète pour y faire des recherches et apporter des idées nouvelles; ils retournent ensuite à leur institution, mieux informés sur les organismes fédéraux et plus au fait des bases de données de l'administration publique et des questions statistiques qui préoccupent les organismes fédéraux. Le National Research Council, qui est une ramification de la National Academy of Sciences, possède un comité des statistiques nationales (Committee on National Statistics) où des statisticiens du secteur universitaire et du secteur privé rencontrent des représentants des organismes gouvernementaux. Réunis en panel ou dans des sessions informelles, ces spécialistes font connaissance et discutent des problèmes les plus courants.

Bien que ces "intermédiaires" et d'autres du même genre ne suffiront pas à faire disparaître la démarcation entre les secteurs, nous croyons que leur présence aura un effet déterminant sur le développement des méthodes d'enquête. Grâce à ces intermédiaires, les progrès réalisés dans un secteur sont communiqués plus rapidement aux autres secteurs mais ce qui est peut-être plus important encore, c'est que les problèmes que doit résoudre un secteur en particulier font éventuellement l'objet de recherches dans tous les secteurs.

REMERCIEMENTS

La rédaction de cet article a été rendue possible en partie grâce à une subvention de la National Science Foundation à la Carnegie Mellon University (no SES-8701606) et à la State University of New York de Stony Brook (no SES-8701816). Une première version est parue sous le titre "Some History of Survey Methods and Data Collection Technology" dans le recueil des actes de 1989 de l'American Statistical Association intitulé *Sesquicentennial Invited Paper Sessions*, 393-405.

BIBLIOGRAPHIE

- ABOWD, J.M., et ZELLNER, A. (1985) Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- ANDERSON, M.J. (1988). *The American Census. A Social History*. New Haven: Yale University Press.
- AMERICAN ECONOMIC ASSOCIATION (1899). *The Federal Census*. Report of the committee on the Twelfth Census. *Publications of the American Economic Association*, New Series, n° 2, 1-7.
- BAILLAR, B.A. (1990). Contributions to statistical methodology from the federal government. *Techniques d'enquête*, ce numéro.
- BOOTH, C. et coll. (1889-1891). 1902-1903. *Life and Labours of the People in London*. London: Macmillan.
- BORUCH, R.F., et PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.

statistiques de l'emploi sur les aspects cognitifs des enquêtes touchant la main-d'oeuvre; le Groupe de travail s'arrête notamment à l'expression "cherche du travail" – problème épineux au sein d'une culture et encore plus complexe entre cultures (Schwarz 1987); notons aussi les travaux qui proposent une double perspective (cognitive et statistique) pour l'analyse de questions comme l'intégration d'expériences dans des enquêtes (Fienberg et Tanur 1989); mentionnons enfin les conférences internationales qui ont porté sur la frontière entre la cognition et les méthodes d'enquête (voir, par exemple, Hippler, Schwarz et Sudman 1987). Pendant que les techniques de la psychologie cognitive servent à la conception des questionnaires, les résultats des tests faits dans les laboratoires de psychologie cognitive sont soumis à des essais pratiques pour que l'on puisse en évaluer l'utilité pour les sociétés de sondage et vérifier s'ils sont généralisables, ce qui contribue à l'enrichissement de la discipline. Notons ici un autre cas d'interaction entre le monde universitaire et l'administration publique: selon des constatations faites en laboratoire, les gens se rappellent plus facilement et plus clairement les visites qu'ils ont faites chez des spécialistes de la santé s'ils commencent par la toute première (Fathi, Schooler et Loftus 1984). Une étude est en cours actuellement pour vérifier si cela peut être observé effectivement dans le cadre de la NHIS (White et Berk 1987).

La tendance visant à faire intervenir les méthodes des sciences cognitives dans la conception des plans de sondage est importante pour plusieurs raisons. Premièrement, elle permet de poser le problème de la préparation des questionnaires dans une perspective scientifique nouvelle. Deuxièmement, elle associe le domaine des sondages à l'étude de phénomènes cognitifs particuliers. Mais ce qui est plus important encore, c'est qu'elle a donné un nouvel élan aux entreprises de sondage et soulevé à nouveau des questions touchant la structure de l'interview, qui vont bien au-delà de la conception des questionnaires et que de nombreux statisticiens croyaient résolues dans les années 1940 et 1950.

6. REMARQUES

Les ouvrages qui relatent l'histoire des méthodes d'enquête insistent habituellement sur le rôle des méthodes d'échantillonnage probabiliste et les perfectionnements qui leur ont été appliqués au fil des ans et, parfois, sur l'étude des erreurs non dues à l'échantillonnage. Dans cet article, nous avons voulu exposer cette histoire dans la perspective des études sociologiques qui se sont faites au cours du 19^e siècle et dans les premières décennies du 20^e siècle et dans la perspective des institutions, publiques ou privées, qui ont contribué au perfectionnement des méthodes d'enquête. Cette approche devrait rappeler au lecteur que les progrès de la théorie statistique ne sont pas la seule chose qui a déterminé l'évolution de la théorie des sondages jusqu'à aujourd'hui. Elle devrait aussi lui permettre de suivre l'évolution de la théorie et de la pratique des sondages à mesure que celles-ci sont influencées par l'évolution des institutions. Il y a un autre aspect de la question que nous n'avons pas encore abordé jusqu'à maintenant. Nous avons dit plus haut que la démarcation entre les trois secteurs ou groupes d'institutions (administrations publiques, sociétés d'études de marché et de sondage (entreprise privée) et universités et autres établissements d'enseignement supérieur) n'était pas absolue. Nous croyons que cette démarcation est de moins en moins réelle à cause de la présence de plus en plus forte d'un quatrième type d'institution que nous appellerons les "intermédiaires". Nous avons vu plus haut comment le Comité on Government Statistics and Information Services de l'ASA et du SSRC, qui se voulait un lien entre le monde universitaire et l'administration publique, a jeté les bases d'un organisme fédéral de coordination des activités statistiques. L'ASA et le SSRC continuent d'agir comme intermédiaires mais il existe aussi d'autres institutions de ce genre. Quelques exemples frappants nous viennent à l'esprit. Ainsi, depuis plus de 40 ans l'American Association for Public Opinion Research réunit des représentants des trois groupes d'institutions dans des sections locales et des conférences nationales où l'on fait état des dernières

Les idées reçues sur la conception et l'analyse des enquêtes longitudinales et à s'interroger sur la définition d'une famille longitudinale (pour une analyse, voir Fienberg et Tanur 1986). Dans les années 1980, on s'est intéressé encore davantage aux enquêtes longitudinales et on a prêté une attention particulière à certains aspects de l'erreur non due à l'échantillonnage, comme le phénomène d'attrition, et à des questions touchant la gestion et l'analyse des données. L'ouvrage de Kalton et coll. (1989) renferme un certain nombre d'articles sur ces sujets.

5.3 Aspects cognitifs des enquêtes

Par les recherches systématiques auxquelles ils se livrent depuis une quarantaine d'années en vue d'améliorer les méthodes d'enquête, les sondeurs ont réussi à mettre au point des méthodes d'interview et de conception de questionnaires très perfectionnées afin de réduire les erreurs non dues à l'échantillonnage, comme celles qu'énumère Deming (voir, par exemple, Payne 1951), et ils ont réalisé de nombreuses études scientifiques pour tester certains aspects de ces méthodes (voir Sudman et Bradburn 1974, Bradburn et Sudman 1979 et Sudman et Presser 1981). Jusqu'à récemment, la recherche visant à approfondir le processus d'interview était peu systématique. Les sondeurs ont commencé à s'intéresser de plus près au processus d'interview lorsqu'ils ont réalisé qu'ils pouvaient recourir à d'autres disciplines, notamment la psychologie cognitive, pour approfondir leur sujet.

Parmi les erreurs non dues à l'échantillonnage, il y a celles engendrées par les processus cognitifs que doivent mettre en jeu le répondant et l'intervieweur durant une interview. Le répondant doit souvent se rappeler des événements et exercer son jugement et doit toujours évaluer la portée des questions qui lui sont posées – la signification de ces questions pour le répondant et pour l'intervieweur. Les sondeurs commencent à peine à tirer profit des notions de psychologie cognitive et de l'expertise des spécialistes du domaine pour analyser plus systématiquement les erreurs non dues à l'échantillonnage. Il convient de souligner que l'approfondissement de la signification n'est pas une activité nouvelle pour les sociétés de sondage. En effet, Cantril (1944) consacre deux chapitres à la présentation des résultats d'expériences qu'il a menées sur la signification et le libellé des questions. Dans ces expériences, il avait utilisé bon nombre des méthodes d'approfondissement et des méthodes de génération de paraphrases qui sont utilisées aujourd'hui dans les laboratoires de psychologie cognitive.

C'est en 1981 que l'on a commencé à vouloir analyser les aspects cognitifs des enquêtes; en effet, le Bureau of Social Science Research et le Bureau of Justice Statistics avaient organisé une conférence où des psychologues de la cognition et des spécialistes des sondages étaient invités à se pencher sur la National Crime Survey. En 1983, le Committee on National Statistics (NSTAT) du National Research Council organisait une conférence plus spécialisée où il invitait les deux groupes de spécialistes à se pencher sur la National Health Interview Survey (Jabine et coll. 1984). Dès l'origine, cette "entreprise" était destinée à réunir des représentants des universités, des instituts de recherche et d'autres établissements d'enseignement supérieur, et de l'administration publique.

Une conséquence directe de la conférence du NSTAT a été l'inauguration du Questionnaire Design Research Laboratory au U.S. National Center for Health Statistics; ce laboratoire, qui est sous la direction de Monroe Sirken, a pour mission de tester au préalable et de roder les principales enquêtes réalisées par l'Etat. Son personnel se compose de fonctionnaires et de chercheurs invités; il accorde des contrats à des universitaires et à des membres d'instituts de recherche en vue de mener à bien sa mission. Des laboratoires semblables ont été ouverts par la suite au Bureau of Labor Statistics et au Bureau of the Census. Une autre conséquence de la conférence du NSTAT a été la création du Committee on Cognition and Survey Research du Social Science Research Council; il s'agit d'un comité multidisciplinaire auquel siègent des représentants de divers établissements. Ce comité a favorisé la recherche sur des sujets comme le processus interactif de l'interview, les utilisations et les pièges de la mémoire rétrospective et l'évaluation de la souffrance dans une enquête. Parmi les autres conséquences de ce mouvement d'exploration notons l'analyse faite par le Groupe de travail de l'OCDE sur les

valoir surtout leur utilité pour la documentation et la normalisation et leur souplesse pour l'interviewer. Bien que les organismes gouvernementaux aient manifesté très tôt de l'intérêt pour l'ITAO, ce n'est que tout récemment qu'ils ont commencé à utiliser de tels systèmes, parfois à titre expérimental et souvent concurremment avec d'autres méthodes de collecte de données; par exemple dans les enquêtes par panel, où la première interview est faite sur place. À ce moment-ci, nous assistons aux débuts de l'interview sur place assistée par ordinateur (IPAO), rendue possible grâce aux recherches qui ont permis de mettre au point les ordinateurs de giron.

5.2 Enquêtes longitudinales

Bien que des enquêtes par panel aient été réalisées lors des campagnes présidentielles de 1924 et de 1940 aux États-Unis (Rice 1928; Lazarsfeld et coll. 1944), il faut attendre les années 1960 avant de voir les spécialistes des études sociologiques s'intéresser vraiment aux données longitudinales. Cela est dû autant plus surprenant que la Current Population Survey a toujours été caractérisée par un plan avec renouvellement et que depuis 1953, de nombreux répondants sont interviewés jusqu'à 8 fois dans un intervalle de 16 mois. À l'origine, on avait doté la CPS d'un plan avec renouvellement dans le but explicite d'obtenir des estimations de la variation d'agré-gats plus efficaces que celles établies à l'aide de données transversales; néanmoins, on aurait pu théoriquement appliquer la structure par panel à la CPS dans son ensemble. Le fait que l'échantillon de cette enquête soit un échantillon d'adresses et non de personnes ou de ménages rend impensable l'utilisation de la CPS sous forme d'enquête par panel (voir les commentaires pertinents sur la National Crime Survey dans Fienberg 1978) mais n'en empêche pas l'utilisation intensive dans l'analyse des flux bruts de la main-d'œuvre (voir, par exemple, Abowd et Zellner 1985 et Stasny 1988).

Il n'est pas nécessaire que tous les sondages visant à mesurer des variations reposent sur des données longitudinales; souvent, des données transversales permettront de mesurer avec au moins autant d'efficacité la variation d'agré-gats. Dans les années 1970, la maison Gallup et d'autres avaient déjà pris l'habitude de poser les mêmes questions d'une fois à l'autre et de publier les résultats dans les journaux. Ces séries chronologiques s'inscrivaient dans la vague naissante des indicateurs sociaux. En 1972, le National Opinion Research Center a mis en oeuvre la General Social Survey (GSS), qui était parrainée par la National Science Foundation. Conçue par un groupe d'universitaires provenant de divers milieux, la GSS a pour but de produire à intervalle régulier des données sur les indicateurs sociaux et de constituer une série de données originales à l'intention des étudiants et des universitaires qui bénéficient de petites subventions de recherche. Afin d'assurer la continuité des données, on a inclus dans la GSS de nombreuses questions qui avaient été formulées à l'origine par Gallup et d'autres maisons de sondage; cette opération fut le prétexte à une fructueuse collaboration entre les établissements (voir, par exemple, Smith 1975).

L'objet fondamental des enquêtes longitudinales est de mesurer des variations dans le temps; toutefois, cela ne se fait pas en comparant les variations de valeurs agré-gées mais les variations de valeurs individuelles. Ce genre d'enquêtes mettent surtout l'accent sur les changements de situation, la durée d'activités et les événements qui se déroulent sur une certaine période. L'intérêt pour les enquêtes longitudinales s'est surtout manifesté, à l'extérieur de l'administration publique; mentionnons par exemple la Panel Study of Income Dynamics, qui est réalisée annuellement depuis 1968 par l'Institute for Social Research de l'Université du Michigan; aussi, les National Longitudinal Surveys of Labor Market Experience, qui, à partir de 1966, ont été parrainées par le Center for Human Resources Research de l'University of Ohio State et qui sont maintenant financées par le BLS; enfin, la Longitudinal Retirement History Survey, parrainée par la Social Security Administration de 1969 à 1979. Durant les années 1970, l'utilisation des enquêtes longitudinales s'est intensifiée, particulièrement au sein des administrations publiques (voir, par exemple, Boruch et Pearson 1988), mais la méthodologie fondamentale de ces enquêtes ressemblait souvent à celle des enquêtes transversales classiques. Ce n'est qu'à la fin des années 1970 que les spécialistes commencèrent à remettre en question

5. ÉVOLUTION DES MÉTHODES D'ENQUÊTE DEPUIS 1960

Au cours des années 1960 et 1970, les enquêtes et les sondages sont devenus monnaie courante aux États-Unis; ce mouvement s'est amorcé au moment de la campagne présidentielle de 1960, durant laquelle les deux candidats (Kennedy et Nixon), qui se disputaient une chaude lutte, commandaient des sondages auprès de l'électorat afin d'évaluer leur position relative. Dans cette section, nous nous intéressons particulièrement à trois aspects des sondages qui ont subi des changements majeurs dans les dernières décennies. En ce qui concerne des sujets aussi importants que l'imputation (dans le cas de données incomplètes) et l'incassante controverse qui entoure l'inférence statistique, le lecteur est prié de consulter d'autres ouvrages (voir, par exemple, Fienberg et Tanur 1983, 1986).

5.1 Mode d'interview: le rôle du téléphone et de l'ordinateur dans les enquêtes

Le progrès et la diffusion de la technologie, particulièrement en ce qui a trait à la téléphonie et à l'informatique, ont largement influencé les méthodes d'enquête durant les années 1960 et 1970. En 1936, on estimait à seulement 35% la proportion des ménages qui avaient le téléphone aux États-Unis; cette situation n'était d'ailleurs pas étrangère aux difficultés qu'éprouvait le *Literary Digest* dans ses sondages (Massey 1988). Cette proportion était passée à 75% en 1960 et à 88% en 1970 et devait se situer autour de 93% en 1986 (Thornderry et Massey 1988). C'est ce qui explique que les enquêtes par téléphone, qui reposent souvent sur le sondage téléphonique au hasard, soient devenues de plus en plus fréquentes et de plus en plus précises. L'idée des enquêtes téléphoniques est venue des entreprises de sondage; les organismes statistiques américains et les universités hésitaient à emboîter le pas parce qu'ils craignaient qu'il y eût surreprésentation ou sous-représentation de certains groupes (selon le niveau de revenu ou l'origine raciale par exemple) (Trewin et Lee 1988) et que l'échantillon ne fût pas assez représentatif. De fait, au sein des organismes publics, l'interview téléphonique sert presque essentiellement aux opérations de suivi, le contact initial ayant été fait sur place (c'est le cas notamment de la *Current Population Survey*, où, depuis 1954, on a recouru à l'interview téléphonique dans les derniers mois de l'enquête). Ce n'est que tout récemment que les organismes statistiques se sont mis à utiliser de façon beaucoup plus systématique le sondage téléphonique au hasard. Groves et Kahn (1979) passent en revue les ouvrages qui ont été écrits sur l'interview téléphonique et illustrent d'une façon générale la comparabilité des données d'enquête en comparant des données recueillies au moyen d'interviews sur place à des données recueillies au moyen d'interviews téléphoniques.

Avec l'apparition et la diffusion rapide de l'ordinateur, les tâches liées à l'analyse des données d'enquête allaient désormais être exécutées plus rapidement que jamais et leur champ d'application allait être plus grand que jamais. On assista alors à une augmentation du nombre des enquêtes réalisées par les diverses institutions, publiques ou privées. Rétrospectivement, il paraît tout à fait naturel de combiner la technique de l'ordinateur à celle du téléphone pour obtenir des systèmes d'interview téléphonique assistée par ordinateur (ITAO). Ces systèmes administrèrent des questionnaires automatisés qui déterminaient eux-mêmes l'ordre des questions et font apparaître les questions pertinentes sur un écran, ils déterminaient le moment où se font les appels et les rappels (et composent souvent eux-mêmes le numéro), ils exécutent les randomisations et automatisent l'entrée des données, sans compter d'autres tâches. Les systèmes ITAO ont été mis au point par des sociétés d'études de marché américaines au début des années 1970 dans le but notamment de suivre l'évolution des caractéristiques des répondants et de veiller par conséquent à ce que les quotas soient respectés à tous points de vue (Nicholls 1988). Chilton Research a été l'une des premières entreprises à utiliser l'ITAO; elle s'en est servie pour des enquêtes visant à connaître le degré de satisfaction des consommateurs vis-à-vis des services offerts par les compagnies de téléphone (Nicholls et Groves 1986). De leur côté, les organismes d'enquête des universités ont commencé à élaborer leurs propres systèmes ITAO au milieu des années 1970 et les ont présentés aux autres membres de la communauté statistique en faisant

L'analyse des données d'enquête (voir Fienberg et Tanur, 1987, 1988 pour une analyse pertinente des liens entre échantillonnage et expérimentation au point de vue du plan et de l'analyse). Lors-que les représentants de ces deux écoles se mettent à publier les résultats de leurs recherches respectives dans diverses revues de statistique durant les années 1940, on observe un nombre de plus en plus grand de points communs.

Au cours des années 1940, l'usage des méthodes d'échantillonnage probabiliste s'étend rapidement à d'autres organismes gouvernementaux. Ce n'est toutefois qu'en 1948, année où l'efficacité des sondages électoraux est durement remise en question (Mosteller et coll. 1949) que les sociétés d'études de marché et les autres optent pour l'échantillonnage probabiliste. Aujourd'hui encore, de nombreuses organisations utilisent une méthode d'échantillonnage proba-biliste avec quotas (Sudman 1967).

Dans cette vague d'approfondissement de la théorie et de la pratique des sondages probabilitistes, on s'est aussi intéressé au phénomène de la non-réponse et aux autres formes d'erreur non due à l'échantillonnage. Dans une analyse d'ouvrages portant sur les erreurs commises dans des enquêtes, Deming (1944) relève 13 facteurs qui influent sur l'utilité fondamentale des enquêtes (notons qu'il s'agit pour la plupart d'erreurs non dues à l'échan-tillonnage):

1. variabilité des réponses;
2. différences entre les divers types et les divers degrés de sondages d'opinion;
3. biais et variation imputables à l'intervieweur;
4. biais dû aux organisateurs de l'enquête;
5. erreurs dans la conception du questionnaire et des schémas de totalisation;
6. modification de l'univers avant publication des données;
7. biais dû à la non-réponse (y compris les cas d'omission);
8. biais dû à la remise tardive de questionnaires;
9. biais dû à un choix peu judicieux de la date de l'enquête ou de la période visée;
10. biais dû à un choix peu judicieux des répondants;
11. erreurs et biais dus à l'échantillonnage;
12. erreurs de traitement (codage, contrôle, calcul, totalisation, pointage, etc.);
13. erreurs d'interprétation.

La plupart des erreurs énumérées par Deming avaient, à cette époque, déjà fait l'objet de recher-ches ou allaient éventuellement en faire l'objet au Bureau of the Census.

Afin de mieux comprendre et de modéliser les erreurs dues à la non-réponse, on a élaboré, dans le cadre du programme de planification et d'évaluation du recensement de 1950 (Hansen, Hurwitz, Marks et Mauldin 1951), un modèle intégré pour les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage dans les recensements et les enquêtes; cette initiative allait s'avérer une étape importante dans le développement des connaissances en statistique. Ce modèle, qui s'apparente à une analyse de variance, ou des variantes de celui-ci ont servi de base à la plupart des recherches qui ont été faites sur l'erreur non due à l'échantillonnage au cours des 35 dernières années tant à l'intérieur qu'à l'extérieur du Bureau of the Census. Brooks et Bailar (1978) font une excellente analyse qualitative de la structure d'erreurs de la Current Population Survey tandis que Mosteller (1978) et Fienberg et Tanur (1983) passent en revue les ouvrages portant sur les erreurs non dues à l'échantillonnage. Enfin, notons que dans son dernier ouvrage, Groves (1989) présente sous un éclairage nouveau une variante du modèle mentionné ci-dessus en ayant soin de faire une distinction entre les éléments aléatoires et les éléments fixes qui découlent des diverses sources d'erreur.

L'article de Bailar (1990) renferme une analyse détaillée sur les erreurs non dues à l'échan-tillonnage dans la perspective du Bureau of the Census.

recherche affiliées. Lazarsfeld était entré aux Etats-Unis en 1933 avec la ferme intention de donner un caractère scientifique aux méthodes qui avaient été élaborées pour les études de marché. Il avait alors créé le Office of Radio Research, qui allait s'appeler plus tard le Bureau of Applied Social Research, à l'université Columbia. Parmi ses nombreuses réalisations, notons l'utilisation de panels et la création d'un système d'analyse causale.

Hadley Cantril était un universitaire qui très tôt collabora avec Lazarsfeld à l'étude des cotes d'écoute de la radio. Lorsque les deux chercheurs décidèrent de poursuivre leurs recherches chacun de leur côté, Cantril créa le Office of Public Opinion Research à l'université Princeton. On y réalisait des études visant à améliorer les méthodes de collecte de données. Par exemple, en analysant les effets du libellé des questions, Rugg et Cantril (1944) ont observé que dans un laps de 6 semaines en 1940 et 1941, le pourcentage d'Américains qui étaient d'accord pour "apporter de l'aide [à la Grande-Bretagne] même au risque d'entrer en guerre" variait de 56 à 78%. En même temps, le pourcentage d'Américains qui étaient d'accord pour que leur pays "entre en guerre immédiatement" variait de 8 à 22%.

Renais Likert a commencé par enseigner à l'université de New York tout en participant à l'élaboration des enquêtes de la Psychological Corporation. Il a ensuite fondé sa propre entreprise, où il réalisait une enquête sur l'attitude des agents d'assurance-vie, dans laquelle il comparait des méthodes qualitatives et quantitatives (surtout des questionnaires). Par la suite, il est devenu directeur de la Division of Program Surveys du Département de l'agriculture. A ce titre, il s'est surtout appliqué à uniformiser les questionnaires. Lorsque Likert quitta le Département de l'agriculture après la Seconde Guerre mondiale, il se dirigea avec son groupe vers l'université du Michigan, où il créa le Survey Research Center.

4. DE LA THEORIE DES SONDAGES A L'ETUDE DES ERREURS NON DUES A L'ECHANTILLONNAGE

Nous avons vu plus haut que les organismes statistiques américains avaient commencé à utiliser l'échantillonnage probabiliste à une époque où on observait des progrès notables dans de nombreux domaines de la statistique et où on jetait les bases d'un processus d'expérimentation et d'inférence sous la direction de statisticiens comme R.A. Fisher, Walter Shewart, Jerzy Neyman et Egon Pearson. Parmi ceux qui ont contribué à la réalisation du recensement d'essai des chômeurs au Bureau of the Census, notons Calvert Dedrick, Morris Hansen, Samuel Stouffer et Frederick Stephan (Anderson 1988; Duncan et Shelton 1978). On a ensuite demandé à Hansen et à quelques autres spécialistes d'examiner d'autres possibilités d'application de l'échantillon du recensement des chômeurs de 1937. Après avoir travaillé à l'échantillon du recensement décennal de 1940 (sous la direction de Deming), Hansen s'est appliqué avec d'autres (notamment, Jerome Cornfield, Lester Frankel, William Hurwitz et J. Steven Stock) à remanier l'enquête sur les chômeurs en se fondant sur de nouvelles idées relatives à l'échantillonnage probabiliste à plusieurs degrés et à l'échantillonnage en grappes (Hansen et Hurwitz 1942, 1943). Le groupe de chercheurs a élaboré une méthode qu'il a ensuite expérimentée dans diverses enquêtes du Bureau of the Census, souvent en collaboration avec d'autres statisticiens. Ces travaux ont abouti en 1953 à la publication d'un compendium en deux volumes de théories et de méthodes (Hansen, Hurwitz et Madow 1953). L'interview accordée il y a quelques années par Hansen (Oikarinen 1987) et l'ouvrage de Duncan et Shelton (1978) nous fournissent des renseignements intéressants et détaillés sur les événements qui se sont déroulés durant cette période.

Outre les travaux énumérés ci-dessus, il convient de souligner les recherches faites par P.C. Mahalanobis et des étudiants de l'Inde et celles faites par Frank Yates et William Cochran d'Angleterre sur l'échantillonnage statistique en agriculture. L'article de Cochran (1939) mérite particulièrement notre attention parce qu'il propose l'utilisation de l'analyse de variance dans l'échantillonnage et qu'il introduit les notions de superpopulation et de modélisation dans

s'occupaient de faire des sondages électoraux et d'en publier les résultats. À l'époque, comme aujourd'hui d'ailleurs, ce genre de sondages était considéré comme l'enquête la plus concluante qui soit. On présupait qu'une organisation qui était réputée pour prévoir avec assez d'exactitude les résultats d'élections pouvait se prononcer avec autant d'exactitude sur d'autres questions, plus difficilement vérifiables.

En ce qui concerne les études de marché, un mouvement semblable se dessine à la toute fin du 19^e siècle; on cherche alors à déterminer les goûts des consommateurs et à évaluer l'effet de la publicité. De là à mesurer l'opinion du public sur d'autres questions de nature concrète ou abstraite, il n'y avait qu'un pas. Vers le milieu des années 1930, on comptait déjà plusieurs sociétés d'études de marché bien établies. Bon nombre d'entre elles menèrent des sondages électoraux en 1936 et obtinrent des résultats beaucoup plus précis que le Literary Digest. Ce sont les dirigeants de ces sociétés (par ex.: Archibald Crossley, George Gallup et Elmo Roper) qui ont contribué à populariser les sondages – sondage électoral, sondage d'opinion et sondage des habitudes de consommation – à la veille de la Seconde Guerre mondiale.

À cette époque, les méthodes de collecte de données connurent un développement remarquable au sein des sociétés d'études de marché et de sondage. L'échantillonnage se faisait soit par choix raisonné ou par la méthode des quotas. La taille des échantillons était élevée et on augmentait progressivement cette taille jusqu'à ce que la loi des grands nombres fasse en sorte que la valeur estimée de la moyenne ou de la proportion se stabilise. Certains questionnaires n'avaient rien de catégorique, l'intervéu n'étant tenu qu'à quelques questions obligatoires – nous appellerions cela aujourd'hui une interview non structurée. D'autres questionnaires étaient plus formels, mais aussi plus courts. Pour expliquer les progrès qui ont été réalisés à l'époque, nous pourrions dire qu'à partir du moment où le niveau de scolarité, le degré de formation et le nombre même des interviewés se sont accrus et que ceux-ci se sont mis à avoir une meilleure connaissance des projets de recherche, les interviews sont devenues plus formelles.

Les enquêteurs de l'époque se préoccupaient des mêmes choses que les enquêteurs d'aujourd'hui. Quelle devrait être la proportion de questions ouvertes et de questions fermées dans le questionnaire? (L'usage semble avoir favorisé une combinaison des deux; le Literary Digest fut le premier, en 1925, à imaginer un "thermomètre d'opinion" pour calibrer les réponses.) En ce qui a trait à la façon de poser des questions délicates – sur l'âge, le revenu, la profession et le mode d'occupation d'un logement, les sondateurs ont résolu le problème en utilisant des listes de contrôle fondées sur le même principe d'opération que les moyens visuels d'aujourd'hui. Par ailleurs, les maisons de sondage se livraient à des expériences portant sur le libellé des questions.

À leurs débuts, comme aujourd'hui d'ailleurs, les sociétés d'études de marché accordaient nécessairement la plus grande importance à l'actualité des résultats. Cela avait le plus souvent pour effet, comme cela se produit encore aujourd'hui, de créer des différends entre les universitaires et les responsables d'études de marché, les premiers croyant que les seconds étaient uniquement motivés par l'argent et ne se préoccupaient guère des règles fondamentales de la discipline, et les seconds croyant que les premiers étaient absorbés par les questions d'ordre abstrait. Il convient toutefois de souligner qu'une des premières maisons de sondage d'opinion et d'études de marché fut la Psychological Corporation, un regroupement de psychologues universitaires désireux de réinvestir une partie de leurs bénéfices dans la recherche. La Psychological Corporation réalisait ses enquêtes par l'intermédiaire de sa Division des études de marché, qui avait été mise sur pied et était dirigée par Henry C. Link.

3.4 Universités

Les universités n'étaient pas totalement absentes du monde des sondages. Dès 1911, la Harvard Graduate School of Business avait créé le Bureau of Business Research pour réaliser des études de consommateurs. Des spécialistes des sciences sociales aussi connus que Paul Lazarsfeld, Hadley Cantril et Rensis Likert se joignirent à des universités et à des instituts de

3.2 Le comité mixte de l'ASA et du SSRC et l'institutionnalisation de l'échantillonnage probabiliste: l'amorce d'une transformation

Au début de la Grande crise des années 1930, les organismes statistiques américains ont de la difficulté à répondre à la demande de statistiques qui vise à surveiller les effets des programmes du New Deal du président Franklin Roosevelt. En 1933, le secrétaire du Travail, Frances Perkins, demande à Stuart A. Rice, président de l'ASA, de mettre sur pied un comité consultatif sur les programmes du BLS. Ce comité allait devenir le Committee on Government Statistics and Information Services (COGIS), une initiative de l'ASA et du Social Science Research Council (SSRC). Duncan et Shelton (1978) donnent un compte rendu détaillé des activités du COGIS; pour les besoins de notre présentation, nous nous arrêtons à deux points en particulier.

Premièrement, en 1933 le COGIS recommanda la création d'un bureau central de la statistique (Central Statistics Board - CSB) pour coordonner les activités statistiques de l'Etat. Une fois qu'ils eurent jeté les bases d'un système statistique fédéral intégré, le COGIS et le CSB entreprirent au début de 1934 de favoriser un accord entre le Bureau of the Census et le BLS, en vertu duquel le premier recueillerait des données de base sur la production et la main-d'oeuvre pour le second.

En deuxième lieu, le COGIS promut l'utilisation des méthodes d'échantillonnage probabiliste dans divers secteurs de l'administration fédérale et incita les employés des organismes statistiques à faire des recherches sur la théorie des sondages. Par exemple, afin d'instituer un cadre technique pour les estimations du niveau de chômage, le COGIS et le CSB organisèrent un recensement d'essai dans le cadre du programme de travaux publics de l'administration fédérale; ce recensement, qui avait pour thème le chômage et qui reposait sur des méthodes d'échantillonnage probabiliste, fut réalisé dans trois municipalités particulières à la fin de 1933 et au début de 1934. Les résultats positifs de ce recensement expérimental et l'accord inter-organismes évoqué plus haut ont amené en 1940 la création de la première grande enquête par sondage permanente ayant pour objet l'emploi et le chômage et reposant sur des méthodes d'échantillonnage probabiliste. Cette enquête allait devenir plus tard la Current Population Survey.

La croisade du COGIS pour l'échantillonnage probabiliste se répécuta à l'Ecole d'études supérieures du Département de l'Agriculture, où W. Edwards Deming avait invité Jerzy Neyman (1938) à faire un cours sur les méthodes d'échantillonnage et d'autres méthodes statistiques en 1937. Ces cours allaient influencer profondément l'évolution de la théorie des sondages au sein de l'administration publique et des universités.

Nous assistons à cette époque à la convergence de plusieurs facteurs qui tendent à favoriser pour la première fois l'utilisation et le perfectionnement de méthodes d'échantillonnage au sein des organismes statistiques américains. L'existence même de ces organismes était la condition préalable essentielle. Une seconde condition était la publication de l'article historique de Neyman (1934), qui proposait une vision tout à fait nouvelle des méthodes de sondage. Pour enclencher tout le processus de changement, il ne manquait plus que la Grande crise, une nouvelle administration constamment préoccupée d'obtenir des données de qualité pour évaluer l'effet de ses programmes sociaux, et le comité mixte de l'ASA et du SSRC (COGIS).

3.3 Sociétés d'études de marché et de sondage

L'origine institutionnelle des études de marché et des sondages aux E.-U. remonte à tout le moins au début du 19^e siècle, lorsque les journaux enregistrèrent les résultats de votes d'essai. Toutefois, cet exercice visait plus souvent à procurer de la publicité au journal ou à faire augmenter son tirage qu'à faire des prévisions exactes. Converse (1987) relève néanmoins quelques cas où l'exercice présente un peu plus de sérieux; en effet, des magazines aussi réputés que le *Literary Digest* (qui était reconnu pour la précision de ses recherches avant le désastre de 1936)

publique et des agents de recensement. L'Université de Pennsylvanie fut l'une des premières universités à s'engager dans la voie des études sociologiques en embauchant W.E.B. DuBois en 1899 pour qu'il réalise une étude sur les Noirs de Philadelphie; son enquête allait se faire par des visites à domicile. À partir des années 1930, et plus particulièrement dans l'après-guerre, les méthodes d'enquête connurent un développement remarquable dans les trois grandes bases institutionnelles: sociétés d'études de marché, universités et administrations publiques. Mais avant de décrire cet essor, nous allons revenir un peu en arrière pour parler de la création des organismes statistiques américains.

Jean Converse (1987) a rédigé récemment un ouvrage très érudit et très soigné sur l'apparition des sondages aux États-Unis en mettant l'accent sur les sociétés d'études de marché et de sondage et les universités. Notre présentation ressemble étroitement à la sienne. Nous avons distingué les bases institutionnelles pour tenir compte d'une réalité sociale et structurer notre exposé. Cependant, nous aimerions que le lecteur garde présent à l'esprit une autre réalité sociale, à savoir que la démarcation entre les institutions n'est pas absolue. En effet, non seulement il se fait un échange d'idées et de méthodes entre les trois groupes d'institutions, mais aussi, dans une mesure, un échange de personnes puisque des individus changent de secteur au cours de leur carrière.

3.1 La création des organismes statistiques américains

L'histoire des organismes statistiques américains débute en 1863, lorsque le Département de l'Agriculture, qui vient tout juste d'être créé, publie le premier rapport sur les cultures et le bétail dans le but de faire état des approvisionnements alimentaires de l'Union pendant la Guerre civile. Ce rapport reposait sur des données recueillies parmi un échantillon de 2,000 agriculteurs répartis dans 22 États et choisis à l'aide d'un sondage par choix raisonné. Depuis ce temps, le Département de l'Agriculture produit régulièrement des statistiques agricoles; aujourd'hui, cette opération est sous la responsabilité du National Agricultural Statistical Service. À la fin des années 1920, les spécialistes des statistiques agricoles étaient rompus aux exercices de corrélation et de régression (Duncan et Shelton 1978). En 1884, le Congrès vota la création du Bureau of Labor (qui allait devenir plus tard le Bureau of Labor Statistics, BLS), qui avait pour mandat de recueillir des données sur le revenu et les conditions de travail des hommes et des femmes. Sous la direction de Carroll Wright, le premier commissaire, le BLS élargit son champ d'activités et s'intéressa à des questions comme les dépressions, les grèves et les lock-out, les salaires des femmes, le mariage et le divorce et le commerce des spiritueux (Norwood et Early 1984). Avec la création du Bureau of the Census en 1902, les États-Unis peuvent compter dès lors sur trois grands organismes qui ont chacun pour mandat de recueillir périodiquement des données nationales. Au cours des trois premières décennies du 20^e siècle, le rôle de ces organismes s'accrut considérablement et au moment du krach d'octobre 1929, on disposait de données sur divers aspects de la vie économique et sociale. Toutefois en 1932, les méthodes d'échantillonnage probabiliste étaient encore peu courantes au sein de l'administration fédérale (Duncan et Shelton 1978).

Aussi difficilement concevable que cela puisse être de nos jours, où des données fiables sont produites mensuellement sur le chômage, il n'existait pas de données semblables dans les années 1920 et au début des années 1930. Hormis certaines données mensuelles recueillies par le BLS auprès de la plupart des entreprises manufacturières et de quelques entreprises non manufacturières, il n'existait pas de données nationales sur le chômage. Pour le recensement de 1920, on supprima la question qui portait sur le chômage parce qu'on craignait qu'elle ne produisît des données plus ou moins exactes. Elle fut incluse de nouveau dans le questionnaire du recensement de 1930 à cause des nombreuses inquiétudes que soulevait alors la question de l'emploi. La vive controverse qu'ont suscitée les données de 1930 sur le chômage (Van Kleeck 1930) et celles du recensement spécial de janvier 1931 fut particulièrement acrimonieuse (Anderson 1988) et eut des répercussions dans la campagne présidentielle de 1932.

recensement à l'autre (American Economic Association 1899). Il aura fallu 12 recensements permanents. Dans l'intervalle, il y avait eu un accroissement soutenu du nombre de recensements de toutes sortes, dont l'objet ne se limitait pas au simple dénombrement.

2.3 Congrès internationaux de statistique

La transition du recensement à l'enquête par sondage fut lente et laborieuse. Kruskal et Mosteller (1980) relatent quelques aspects de cette transition, plus particulièrement en ce qui a trait aux discussions qui ont eu lieu lors des sessions de l'Institut international de statistique (IIS) au sujet des sondages, et nous nous inspirons largement de leur ouvrage dans cette section. Quetelet a jeté les bases de ces sessions internationales au milieu du XIX^e siècle lorsqu'il participa à l'organisation du premier d'une série de congrès internationaux de statistique en 1853. Après neuf congrès, qui se sont déroulés entre 1853 et 1876, l'IIS a vu le jour en 1885. Il est intéressant de constater que l'index du livre de Stigler (1986) sur l'histoire de la statistique avant 1900 ne contient qu'un ouvrage sur les enquêtes par sondage, soit une étude de Quetelet datant de 1830 et ayant un rapport avec une méthode de recensement proposée par Laplace, et que l'index du livre de Porter (1986) n'en contient que deux, soit une étude de Karl Pearson datant de 1900 et l'ouvrage de Kiaer et de l'IIS.

Déjà à la session de l'IIS de 1895, Kiaer (1895-1896) se prononçait en faveur d'une "méthode représentative" ou d'une "enquête partielle", où l'enquêteur choisirait tout d'abord des circonscriptions, des villes, etc., puis choisirait à l'intérieur de celles-ci des unités (personnes). On procéderait par choix raisonné à chaque niveau en veillant à ce que tous les genres d'unités soient représentés. Si ce principe était respecté et que la taille de l'échantillon à chaque degré d'échantillonnage était élevée, on était d'avis que l'échantillon obtenu était représentatif.

L'idée de réaliser des sondages plutôt que des recensements fut fortement contestée mais Kiaer présenta des arguments en faveur des sondages (arguments appuyés par certains et réfutés par d'autres) aux sessions de l'IIS de 1897, 1901 et 1903. Vers la fin de cette période, on commença à parler d'échantillonnage probabiliste mais d'après les comptes rendus des sessions de l'IIS, la question de la méthode représentative n'aurait pas été réexaminée avant 1925. À cette époque toutefois, la méthode représentative semblait être entrée dans l'usage, si l'on se fonde une fois de plus sur le compte rendu, et les discussions portaient surtout sur la façon d'obtenir un échantillon représentatif et de mesurer la précision des estimations d'échantillon (Bowley 1926; Jensen 1926). À cette même époque, on commença à parler d'échantillonnage en grappes et de stratification mais l'échantillonnage par choix raisonné demeurait la méthode incontestée.

L'efficacité du sondage par choix raisonné sera sérieusement remise en question lorsque Cini et Galvani échantillonneront par choix raisonné des questionnaires d'un recensement fait en Italie et découvriront que les circonscriptions qui ont servi à établir la moyenne nationale pour sept variables ne sont pas assez représentatives pour d'autres variables (Cini 1928; Cini et Galvani 1929). Peu après, Neyman (1934) publiera son article inédit dans lequel il illustre notamment les mérites de l'échantillonnage probabiliste.

3. ÉVOLUTION DES BASES INSTITUTIONNELLES DES SONDAGES AUX ÉTATS-UNIS

Les enquêtes par sondage aux États-Unis sont nées de l'action combinée des trois bases institutionnelles qui ont eu tant d'influence en Europe – particulièrement pour leur propre compte, universités et administrations publiques. Les premières études sociologiques aux États-Unis (avant la Première Guerre mondiale) semblaient être calquées sur le modèle britannique; leur réalisation était confiée à des travailleurs sociaux, des travailleurs du secteur de l'hygiène

de nombreux savants, comme Laplace et Quetelet (un Belge qui était venu étudier en France sous la direction de Laplace), qui, à leur tour, contribuèrent largement au développement de la théorie et de la pratique des enquêtes, se risquant notamment à reprendre des principes de la théorie des probabilités par l'application de ce que l'on appelle aujourd'hui l'estimation par quotient (voir Stigler 1986, Chapitre 5). Après la révolution de 1830, l'Académie des sciences morales et politiques organisa des concours pour inciter les statisticiens à entreprendre des recherches.

En Allemagne, la notion de "statistique" (collecte de données sur l'Etat) était diffusée dans les universités dès la fin du XVIII^e siècle. Au début du XIX^e siècle, on divisa la statistique en trois branches, la statistique descriptive appliquée à la science politique et la statistique historico-mathématique appliquée à l'économie politique alliant demeurer le propre des universités tandis que la statistique des sondages allait être appliquée dans les bureaux de recensement et d'autres organismes d'Etat.

En 1872, on fonda la Verein für Socialpolitik, qui se voulait à la fois un groupe de pression, une association professionnelle et un organisme de recherche. Cet organisme préparait des questionnaires à l'intention de personnes qui étaient réputées bien informées comme les propriétaires, les pasteurs et les notaires. Cependant, à cause du risque d'inexactitude des données fournies par les répondants, de l'ordre aléatoire et de l'imprécision des questions et des taux de réponses faibles, la Verein für Socialpolitik a dû mettre fin à cette activité. Au début du XX^e siècle, Levenstein (1912) publia ce qui fut probablement le premier grand sondage d'opinion, pour lequel il utilisa une méthode boule de neige. A la même époque, Max Weber tenta de réaliser une enquête sur les travailleurs industriels; il prévoyait recueillir une partie de ses données auprès des travailleurs mêmes mais s'aperçut que la majorité d'entre eux refusait de prêter leur concours.

2.2 Les recensements: un prélude aux sondages

Les méthodes d'enquête trouvent aussi leur origine dans l'histoire des méthodes de recensement; nous allons donc faire un court exposé sur les recensements et les infrastructures pertinentes. Beaucoup d'autres auteurs ont observé que le recensement moderne avait son origine dans les recensements décrits dans l'Ancien testament (Madansky 1986) et ceux réalisés par les Egyptiens, les Grecs, les Japonais, les Perses et les Romains dans l'Antiquité (Taeuber 1978). Les écrits bibliques semblent insister plus sur le résultat de ces recensements que sur la méthode de dénombrement bien qu'à plusieurs endroits dans le texte, on évoque la rapidité du processus. Mais pour ce qui nous occupe, transportons-nous directement à la fin du XVIII^e siècle, où a lieu le premier recensement des Etats-Unis d'Amérique; à ce propos, les opinions divergent quant au pays qui aurait été l'initiateur du recensement moderne, à savoir le Canada, la Suède ou les Etats-Unis (Willcox 1930).

Le premier recensement des Etats-Unis a eu lieu en 1790 (le recensement de 1990 coïncidera avec le bicentenaire des recensements aux Etats-Unis); les Etats en avaient alors assuré l'existence et avaient été rembourrés par le gouvernement fédéral. Au recensement suivant, celui de 1800, la tâche avait été confiée aux shérifs adjoints (Duncan et Shelton 1978). Ce n'est qu'en 1880 que le Census Office se vit confier la responsabilité des opérations sur le terrain et se vit autorisé à nommer des recenseurs.

Avant 1850, l'unité de référence dans les recensements aux E.-U. était la famille; on rapportait peu de données sur les personnes mêmes. Le passage de l'optique "familiale" à l'optique "individuelle" est largement attribuable aux travaux de Lemuel Shattuck, l'un des fondateurs de l'ASA (American Statistical Association), qui avait réalisé précédemment le recensement de Boston de 1845 (Anderson 1988, p. 36-37), et à ceux de Quetelet, qui a participé à l'organisation du recensement de 1846 en Belgique (Willcox 1930).

Les méthodes de recensement évoluèrent aux E.-U. puisqu'à tous les 10 ans, on mettait sur pied une organisation qui, conformément à l'obligation constitutionnelle, devait faire le recensement de la population des E.-U.; cependant, il y avait un manque d'uniformité flagrant d'un

Puis il y a eu la crise économique des années 1930, l'éclosion des méthodes d'échantillonnage probabiliste et la création d'un organisme fédéral de coordination des activités statistiques, trois facteurs qui ont marqué le début de la période contemporaine des méthodes d'enquête aux E.-U. Nous analysons également le rôle des sociétés d'études de marché et des universités comme bases institutionnelles. Dans la section 4, nous soulignons le rôle qu'a joué le U.S. Bureau of the Census dans l'étude des erreurs non dues à l'échantillonnage, entreprise dans les années 1940 et 1950. Enfin dans la section 5, nous examinons quelques-uns des principaux changements survenus dans les méthodes d'enquête depuis 1960, notamment en ce qui concerne les progrès techniques, le rôle des enquêtes longitudinales et la récente tendance voulant que l'on examine les aspects cognitifs des enquêtes.

2. SURVOL HISTORIQUE DES FONDEMENTS INSTITUTIONNELS DES MÉTHODES D'ENQUÊTES MODERNES

2.1 Les premières études sociologiques en Europe

Les méthodes d'enquête et de collecte de données qui ont cours aux E.-U. trouvent en partie leur origine dans les premières études sociologiques réalisées en Europe (voir Lecuyer et Oberschall, 1978, dont nous sommes inspirés).

En Angleterre, les premières études sociologiques remontent au XVII^e siècle. Qualifiées d'arithmétique politique, ces études reposaient sur des fichiers administratifs (surtout les registres paroissiaux) et des données recueillies par l'observateur. Elles étaient réalisées le plus souvent par des personnes reconnues pour leur grande rigueur, tel John Graunt, qui a publié ses *Natural and Political Observations Made Upon the Bills of Mortality* en 1662. Jusqu'au début du XVIII^e siècle, la paroisse était le cœur de la vie politique et sociale de sorte qu'il était normal de s'adresser aux membres du clergé pour obtenir des renseignements pour toutes sortes d'enquêtes. Toutefois, la révolution industrielle et la naissance des agglomérations urbaines sont venues mettre fin à cette forme de relation, obligeant désormais les observateurs à aller voir les gens dans leur maison.

Dans les années 1830, des sociétés statistiques étaient créées en Angleterre afin d'étudier des problèmes sociaux. Ces sociétés mettaient sur pied des comités qui recrutèrent des gens pour recueillir des données au domicile des personnes concernées. Bien que les sociétés statistiques étaient dissoutes lorsque les problèmes sociaux semblaient résolus, leurs méthodes ont été reprises vers la fin du XIX^e siècle, lorsque Booth (1889-1891) envoya des agents du conseil scolaire à domicile pour étudier la situation des personnes défavorisées à Londres.

En France, où l'administration était plus centralisée, les premières études sociologiques ont été réalisées par l'Etat. Les intendants de province devaient remplir des questionnaires sur la situation démographique et économique de leur province. Vers le milieu du XVIII^e siècle, l'Etat se livra à ce que l'on pourrait appeler aujourd'hui une étude des effets de la communication de masse. En effet, on avait demandé aux intendants de faire courir des rumeurs de hausse d'impôts et de conscription et de faire rapport sur les réactions du peuple.

Sous Napoléon 1^{er}, le gouvernement français créa un organisme national qui avait pour mission de recueillir des données sur la population, les conditions sociales, l'agriculture et l'activité industrielle et commerciale (Bourguet 1988). Bien que cette initiative n'ait pas produit les résultats escomptés et que les méthodes de recensement utilisées ne fussent pas aussi rigoureuses que celles que nous connaissons aujourd'hui, le gouvernement venait néanmoins de mettre sur pied une structure institutionnelle. Au cours du XIX^e siècle, la France continua d'exercer ses responsabilités à ce chapitre en recueillant des données par l'intermédiaire des préfets et du Bureau de statistique. Sous Napoléon, on avait aussi institué un programme de statistiques sociales qui rejetait explicitement les principes de la théorie des probabilités telle qu'elle était connue à l'époque. L'intérêt que manifestait la France pour les statistiques sociales influença

Origine institutionnelle des enquêtes par sondage aux États-Unis: Une perspective historique

STEPHEN E. FIENBERG et JUDITH M. TANUR¹

RÉSUMÉ

L'idée fondamentale de cet article est que pour bien comprendre l'évolution des méthodes d'enquête au sens technique, il est nécessaire de suivre l'évolution des établissements chargés de la réalisation des enquêtes. C'est pourquoi nous envisageons ici les méthodes d'enquête d'un point de vue général pour que, dans un deuxième temps, le lecteur soit plus à même de saisir la démarche mathématique exposée dans les ouvrages traitant la théorie des sondages. Après une courte introduction, nous résumons l'évolution des facteurs institutionnels et circonstanciels qui ont prévalu en Europe et aux États-Unis jusque dans les premières décennies du XX^e siècle, en insistant sur les administrations publiques. Nous nous intéressons ensuite aux circonstances qui ont entouré la création d'organismes d'enquête aux États-Unis, principalement dans les années 1930 et 1940. Par ailleurs, nous ne manquons pas de souligner le rôle qu'a joué le U.S. Bureau of the Census dans l'étude des erreurs non dues à l'échantillonnage, entreprises dans les années 1940 et 1950. Enfin, nous examinons l'évolution qui s'est faite dans les méthodes d'enquête depuis 1960 à trois chapitres en particulier.

MOTS CLÉS: Recensements; aspects cognitifs des plans de sondage; erreurs non dues à l'échantillonnage; échantillonnage probabiliste; organismes d'enquête.

1. INTRODUCTION

Pour bien comprendre l'évolution des méthodes d'enquête au point de vue technique, il faut nécessairement se référer à l'histoire des établissements chargés de réaliser des enquêtes. Nous nous proposons ici d'envisager les méthodes d'enquête dans cette perspective afin que le lecteur soit plus à même, dans un deuxième temps, de saisir la démarche mathématique exposée dans de nombreux ouvrages sur la théorie des sondages de même que dans l'article de Rao et Bellhousé (1990). Bien que notre approche soit relativement nouvelle, nous avons puisé largement à des sources secondaires qui renferment des présentations détaillées différentes des nôtres. Cet article porte plus spécialement sur l'évolution des méthodes d'enquête aux États-Unis mais il trace aussi un tableau des facteurs historiques qui ont concouru au développement des méthodes d'enquête.

Dans la section 2, nous faisons un bref historique de l'évolution de ces facteurs depuis les origines jusqu'aux premières décennies du XX^e siècle, en faisant référence à deux grands types d'enquêtes: les études sociologiques et les recensements. Nous commençons par un résumé des origines des études sociologiques en Europe, puis abordons brièvement la question des recensements, plus particulièrement en ce qui a trait aux États-Unis, et enfin nous voyons dans quelle mesure les congrès internationaux de statistique de la fin du XIX^e siècle et du début du XX^e ont consacré l'importance de l'échantillonnage. Cependant, ces congrès n'ont pas suffi pour faire comprendre le rôle que pouvait jouer l'élément probabiliste dans les sondages. Pour cela, il fallait d'autres bases institutionnelles.

Dans la section 3, nous soulignons l'apparition de nouveaux organismes d'enquête aux États-Unis, dans les années 1930 et 1940. Déjà au début du siècle, il y avait eu la création des organismes statistiques du gouvernement américain, ce qui venait combler une lacune à l'époque.

¹ Stephen E. Fienberg, College of Humanities and Social Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, E.-U.; Judith M. Tanur, State University of New York, Stony Brook, E.-U.

changent et l'on ne se contente plus maintenant de mentionner les erreurs d'échantillonnage et de donner, en plus, de vagues avertissements à propos de la taille possible des erreurs non dues à l'échantillonnage; on tente plutôt de mesurer l'erreur totale affectant une enquête en reconnaissant que certains des biais non dus à l'échantillonnage peuvent dépasser de beaucoup les erreurs d'échantillonnage.

La section 4 de l'article est consacrée à l'analyse de données d'enquête, aux utilisations analytiques plutôt que descriptives des enquêtes. Ici, la dispute entre l'approche fondée sur un plan et l'approche fondée sur un modèle perd toute importance. Les analystes doivent faire face à tous les problèmes classiques du choix du modèle, de l'estimation et des essais, de l'analyse des résidus et ainsi de suite, qui constituent le courant dominant en matière de statistique. Le sujet négligé qu'est le domaine des enquêtes par sondage est finalement intégré au reste de la statistique.

Mes derniers commentaires sont encore d'ordre personnel. Si vous consultez la bibliographie à la fin de l'article et si vous considérez les domaines additionnels que j'ai mentionnés, vous verrez que Jon Rao a publié des articles importants dans chacun de ces domaines. Je crois qu'il était particulièrement approprié qu'on l'invite à présenter cet article, ses liens étroits avec Hartley et Cochran n'étant pas la moindre des raisons qui justifient ce choix. Je le félicite pour ses contributions aux enquêtes par sondage et je félicite les deux auteurs pour leur excellent article.

BIBLIOGRAPHIE ADDITIONNELLE

- DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society*, séries B, 15, 262-269.
- ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, séries B, 31, 195-233.
- FELLEGI, I. P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SCOTT, A. J. (1977). On the problem of randomization in survey sampling. *Sankhyā*, séries C, 39, 1-9.

d'enquêtes répétées dans un modèle économétrique, je préférerais introduire les estimateurs transversaux avec leur structure de corrélation connue plutôt que des estimateurs "composites". Par contre, si je voulais la meilleure estimation de la valeur actuelle, disons, du chômage, à une fin particulière et que ce renseignement n'était pas destiné au public, alors j'utiliserais la procédure la plus efficace disponible. De même, si je désirais expliquer le changement dans la valeur de certains estimateurs, dans le temps, alors je devrais aller plus loin que la simple analyse effectuée à l'aide de la randomisation. Ainsi, les problèmes relatifs à l'inférence fondée sur la randomisation pour les enquêtes répétées se produisent surtout pour les analyses secondaires. Cependant, il reste la question importante de déterminer quelles estimations devraient être communiquées au public.

La section 2 de l'article est consacrée au travail, sur les fondements théoriques de l'inférence à partir de données d'enquête, effectué au cours des 30 à 40 dernières années. Les auteurs ont choisi de distinguer trois approches: approche fondée sur un plan, approche dépendante d'un modèle et approche fondée sur un modèle. La dernière approche étant une tentative afin de trouver une solution de compromis entre les deux autres. Personnellement, je préfère utiliser une approche globale (théorie universelle) (TU) qui intègre à la fois des plans et des modèles. En plus d'Ericson, Scott (1977) et Rubin (1976) on exerce des influences importantes sur mes opinions dans ce domaine. Dans l'approche globale (TU), les variables d'enquête, le mécanisme de sondage et tous les autres mécanismes de sélection et de mesure sont tous introduits explicitement dans un modèle global. Si Y est la matrice $n \times p$ des variables d'enquête mesurées, z représente les informations préalables, s représente l'échantillon, $s^* \subset s$ représente les répondants, alors la distribution conjointe de toutes ces variables est

$$f(Y | z; \theta) g(z; \phi) p(s | z) q(s^* | s, z, X_s; \eta),$$

où le plan d'enquête, représenté par $p(s | z)$, est du type que l'on dit ne donner aucune information comme l'échantillonnage aléatoire. Le plan ne donne aucune information parce que l'on suppose z (les informations préalables) connues et qu'elles comprennent toutes les informations habituelles sur la stratification, sur l'échantillonnage en grappes et sur les mesures de la taille. La formulation générale force les statisticiens à faire face à toutes leurs hypothèses. La non-réponse doit être modélisée explicitement. Les erreurs de mesure doivent être incluses dans la structure de $f(Y | z; \theta) g(z; \phi)$. La décision d'utiliser l'inférence fondée sur la randomisation est alors une formulation explicite que, étant donné z , l'on peut traiter les valeurs de X comme des constantes inconnues qui sont des valeurs arbitraires à propos desquelles nous n'avons aucune information additionnelle. Par contre, une personne qui utilise un modèle doit définir ce dernier au niveau nécessaire pour l'inférence, par exemple, par un modèle interchangé-ble. L'approche fondée sur un plan ainsi que l'approche dépendante d'un modèle reposent toutes deux sur les mêmes informations préalables, z , et ainsi les deux approches devraient employer des structures semblables et, peut-être, identiques. En fait, je m'attendrais rarement à ce que les estimateurs ponctuels utilisant les diverses approches diffèrent beaucoup en pratique. La question se ramène donc à celle désignée par les auteurs comme le choix d'une mesure de l'incertitude. Les procédures fondées sur un plan sont, à strictement parler, inconditionnelles. La façon de construire des inférences conditionnelles fondées sur un modèle est encore sans réponse, mais l'approche de Robinson (1987) semble prometteuse. Le modèle global (TU) montre que la controverse entre l'approche fondée sur un plan et l'approche fondée sur un modèle est bien ce qu'elle est, c'est-à-dire une dispute philosophique relativement mineure dans le cadre beau-

Il faut considérer que le fait que les statisticiens, tant théoriciens que praticiens, n'ont pu intégrer les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage en des mesures de l'erreur totale affectant une enquête même après 50 ans de recherches intensives constitue l'un des échecs de cette branche importante de la statistique. Mais, encore une fois, les choses

l'interchangeabilité. Royall (1970, 1973) a cependant commis l'erreur de se faire le champion de l'échantillonnage raisonné dans son approche fondée sur un modèle. Il a biqué certaines personnes au vif et s'est attiré la colère de ceux qui font la loi en matière de randomisation. Je pensais que Royall avait posé certaines questions sérieuses qui méritaient une réponse et la vigueur de la réaction m'a surpris. Pourquoi les spécialistes en enquêtes par sondage des universités et ceux des organismes gouvernementaux de l'Amérique du Nord avaient-ils une opinion si arrêtée à propos de la randomisation? Leurs collègues travaillant aux études de marché semblaient heureux avec les échantillons obtenus par la méthode des quotas que l'on pouvait considérer comme un cas spécial de l'échantillonnage équilibré. En Europe, de nombreuses enquêtes officielles sont basées sur des échantillons obtenus par la méthode des quotas. Qu'y a-t-il de si spécial à propos des statistiques officielles en Amérique du Nord?

Je pense que la réponse à cette question se trouve profondément dans le psychisme politique des Américains. Les Américains sérieux sont des démocrates dans le vrai sens du mot. Ils croient à la liberté individuelle et au droit à l'information, ils sont aussi très méfiants des gouvernements. Ils reconnaissent le besoin de disposer, dans une démocratie, de renseignements statistiques fiables. Pour les spécialistes en statistiques officielles, la randomisation est la garantie de la précision objective de leurs données. C'est une source clé de leur intégrité professionnelle et toute attaque contre la randomisation était considérée comme potentiellement dangereuse même si l'intention était très bonne. J'admire cette position et elle a aidé à me convaincre que la randomisation est une des grandes contributions qu'a apporté la statistique à la science.

Je me suis exprimé avec émotion parce que je suis tellement mécontent de la position actuelle des statistiques officielles au Royaume-Uni. La tradition au R.-U. n'est pas naturellement démocratique, nous sommes encore une monarchie, nous respectons l'autorité plutôt que l'individu. Cette tendance est exploitée présentement et il y a actuellement une érosion sérieuse de la confiance du public pour ce qui est de l'utilisation des statistiques par le gouvernement. On a soutenu, qu'au R.-U., les statistiques sont recueillies pour aider les décisions du gouvernement, pas pour aider le parlement ou pour informer l'électorat. On a cessé de recueillir des données pour des séries clés, des définitions ont été modifiées, des informations sont présentées par des ministres de façons manifestement fausses cependant, aucun statisticien gouvernemental ne peut se plaindre publiquement à cause de la loi sur les secrets officiels. On retrouve un cynisme dangereux au sein du public et il se peut que les prédictions faites par George Orwell dans son roman 1984 soient plus près de la réalité que nous ne le croyons. Je m'excuse auprès des auteurs pour cette digression, mais j'ai dit que j'enfouirais mes données et la question de l'intégrité des statistiques officielles a une grande importance.

Avant de laisser la théorie de la randomisation, je désirerais faire quelques commentaires à propos des enquêtes répétées et de l'échantillonnage avec renouvellement. Encore une fois, il s'agit d'un domaine dont les auteurs n'ont pas traité, bien qu'ils aient mentionné Patterson (1950) comme un article jalon. La théorie de la randomisation a été élaborée dans le cadre des enquêtes transversales uniques. Le prolongement de probabilité dans le temps pour l'échantillonnage avec renouvellement quand la population change, Fellegi (1963). Pour la mesure des flux bruts ou des probabilités de transition, le rôle des probabilités de sélection par randomisation n'est pas clair. La belle simplicité de la théorie de la randomisation pour les enquêtes uniques est détruite quand on les répète dans le temps. Mais la majorité des enquêtes importantes sont des enquêtes répétées, particulièrement dans le secteur gouvernemental, quelles sont donc les implications de cette situation?

Comme toujours, la réponse est que cela dépend. Si le but principal est de produire des statistiques descriptives de l'état du système pour chaque période, on peut alors considérer les enquêtes comme des répétitions d'une enquête transversale et chacune d'entre elles peut être analysée indépendamment. Bien que les estimateurs "composites" ou les estimateurs de séries chronologiques puissent être plus efficaces, on devrait les considérer comme des estimateurs secondaires plutôt que comme des estimateurs primaires. Si je désirais utiliser des données

Une troisième raison est le fait que puisque les unités de la population finie peuvent prendre des valeurs arbitraires, la population ne peut être résumée au moyen de quelques paramètres. Des notions telles que l'exhaustivité ont peu de valeur dans la théorie des enquêtes par sondage et les données-échantillons sont habituellement résumées par une masse de tableaux à double entrée. L'estimation d'un grand nombre de proportions pour des cases est le but principal des enquêtes par sondage et l'objet de l'inférence est habituellement descriptif plutôt qu'explicatif. Une dernière raison qui explique pourquoi les enquêtes par sondage sont séparées du courant dominant en matière de statistique est le fait que la théorie de la randomisation appliquée aux enquêtes par sondage est tellement complète. C'est une théorie fermée pour laquelle, si on l'accepte, il reste peu de problèmes à résoudre. Les préoccupations principales des chercheurs qui se sont intéressés à la randomisation, depuis qu'Horvitz et Thompson (1952) ont fourni le modèle théorique général, ont été la construction de plans de sondage ppt avec probabilités de sélection composées non nulles, la production de méthodes et de programmes pour l'estimation de la variance et la construction d'estimateurs qui utilisent des renseignements supplémentaires mais qui ne peuvent jamais être généralement efficaces à cause des résultats de Godambe. Tous ces problèmes sont importants, mais ils ne sont pas passionnants, ils n'ont pas la profondeur philosophique et mathématique nécessaire pour capter l'imagination des jeunes mathématiciens.

Selon moi, ces raisons expliquent pourquoi les enquêtes par sondage ont été considérées, dans la passé, comme une activité reléguée aux limites du courant dominant en matière de statistique. Cette position est en train de changer et je décèle une rencontre des différentes branches de la statistique. Une bonne partie des travaux récents en matière d'enquêtes par sondage ont tenté d'intégrer les enquêtes au courant dominant dans le domaine de la statistique et, dans de nombreux secteurs de la statistique, on reconnaît maintenant l'importance des effets d'échantillonnage. Le sujet négligé qu'est le domaine des enquêtes par sondage serait-il en train d'être intégré au reste de la statistique?

En plus de son théorème qui démontre qu'il ne peut exister de meilleur estimateur non biaisé (pour la catégorie d'estimateurs qu'il a proposés) pour aucun plan d'échantillonnage, Godambe a aussi démontré que dans le modèle de randomisation, la vraisemblance est proportionnelle à la probabilité de sélection, $p(s \mid z)$, où z représente les informations préalables sur lesquelles le plan de sondage était basé, ce qui, pour un s fixe, est une constante. Ainsi, la fonction de vraisemblance ne donne aucune information. Pour le même modèle, Basu (1971) a démontré que la statistique exhaustive est $\{ (i, y_i) : i \in s \}$, normalement la bande de données complète, y compris les labels. Bien que ces résultats soient aussi négatifs, mettant en lumière la distinction entre l'inférence fondée sur la randomisation et les autres formes d'inférence, ils ont stimulé l'intérêt parmi un groupe plus étendu de statisticiens et ils ont ainsi eu une valeur positive. Mon propre intérêt pour la théorie des enquêtes par sondage a été stimulé par Ericson (1969), en particulier par la façon dont il a incorporé la fonction de vraisemblance non informative à un modèle positif au moyen du théorème de Bayes et des distributions a priori interchangeables. L'utilisation qu'a faite Ericson de l'interchangeabilité mérite que tous les statisticiens en tiennent compte, pas seulement ceux de l'école bayésienne. Est-il raisonnable, est-il même possible, d'avoir une théorie valable de l'inférence prédictive sans une certaine forme d'interchangeabilité? S'il n'y a aucune fonction des valeurs unitaires qui est interchangeable, comment peut-on prédire les valeurs qui n'ont pas été observées à partir des valeurs-échantillons? Selon moi, le travail d'Ericson a été un jalon dans l'évolution de la théorie des enquêtes par sondage.

La nature non informative de la fonction de vraisemblance fondée sur la randomisation a amené certains statisticiens à mettre en doute le rôle de la randomisation. Godambe, lui-même, a mentionné «(traduction) le problème de la randomisation» et il a élaboré des approches théoriques de rechangement qui faisaient appel à la randomisation. Ericson aussi a trouvé un rôle pour la randomisation dans son modèle interchangeable. Il a soutenu que si l'on utilise nos informations préalables, z , pour former des groupes d'unités qui sont approximativement interchangeables a priori, alors l'utilisation de l'échantillonnage aléatoire simple garantira

COMMENTAIRES

T.M. FRED SMITH¹

Les enquêtes par sondage sont un des plus importants domaines de la statistique. L'article des professeurs Rao et Bellhouse est un excellent examen de l'évolution de la théorie des enquêtes par sondage et je trouve difficile d'être critique; mais, dans la meilleure tradition de la Royal Statistical Society, je vais tenter de l'être, de façon aussi constructive et controversée que possible. Dans tout rapport de synthèse, le choix des sujets, particulièrement ceux qui se rapportent à des travaux récents, doit être subjectif dans une certaine mesure. Cela offre une cible facile au critique; critiquer les auteurs pour leurs péchés d'omission. L'examen doit être assez étendu et cela permet aux critiques d'enfouir leurs dadas. Je vais adopter les deux approches et, ce faisant, je viserai à relever certaines questions additionnelles que je considère importantes, élargissant ainsi d'avantage l'examen.

Il y a maintenant un accord général à propos des jalons relatifs à notre sujet. Ils sont associés aux noms de Kiaer, Bowley, Neyman, Cochran, Hansen, Hurlitz, Madow, Mahalanobis, Horvitz et Thompson. Une collection internationale dominée, dernièrement, par des contributions des Etats-Unis. Le travail de Kiaer et de Bowley était fondamentalement parce qu'ils ont démontré que l'on pouvait tirer des conclusions valides à partir d'échantillons représentatifs de très petite taille tirés de grandes populations avec des valeurs arbitraires. Les échantillons représentatifs étaient des échantillons stratifiés avec des répartitions proportionnelles et Bowley a obtenu les résultats théoriques appropriés. Neyman et des auteurs qui ont écrit après lui ont défendu la cause de l'échantillonnage aléatoire et élaboré une théorie complète de l'inférence fondée sur la randomisation applicable à la majorité des plans de sondage. Durbin (1953) complète la théorie avec ses résultats sur l'échantillonnage à plusieurs degrés. En dépit de l'importance de ces résultats, les enquêtes par sondage sont devenues un sujet négligé rélégué aux limites du courant dominant en matière de statistique et, même aujourd'hui, la majorité des départements des universités n'emploient pas de statisticien spécialiste en sondages. Pour-quoi en est-il ainsi?

Une raison qui explique cette situation est le fait que la théorie des enquêtes par sondage a été élaborée surtout dans le cadre des sciences sociales et des statistiques gouvernementales officielles, alors que la majorité des statisticiens ont une formation en mathématiques et dans les sciences physiques. Bien que tous les chercheurs oeuvrant dans les sciences expérimentales travaillent avec des échantillons, très peu d'entre eux semblent reconnaître ce fait explicitement et ceux qui le font, comme les géologues et les biologistes, ont élaboré leurs propres méthodes de l'échantillonnage et de l'estimation. Selon moi, il est temps de mettre en contact les experts en sondage de tous les domaines de la recherche scientifique afin qu'ils puissent partager leurs idées et leurs expériences et, espérons-le, établir une théorie globale des enquêtes par sondage. Une deuxième raison est le fait que les enquêtes par sondage commencent avec une population qui est une véritable population finie et de taille fixe d'unités. Des échantillons sont alors tirés de cette population selon des règles précises. Dans la majorité des recherches scientifiques, la situation est inversée; la population n'est pas bien définie et le chercheur commence avec un échantillon. Une vision du rôle du statisticien, énoncée, par exemple, par R.A. Fisher, est de définir la population hypothétique de laquelle on peut considérer les données-échantillons comme un échantillon aléatoire. Cette approche élude la question de savoir si cette population hypothétique a une valeur scientifique quelconque. On pourrait soutenir que beaucoup de choses nous font recommander l'approche utilisée pour les enquêtes par sondage où l'on commence avec la population pour ensuite examiner les rapports entre l'échantillon et la population précisée.

¹ T.M.F. Smith, Département de mathématiques, The University, Southampton, SO9 5NH, R.-U.

SCHNEBL, D., KENNEDY, W. J., SULLIVAN, G., PARK, H. J., et FULLER, W. A. (1988). Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes. *Techniques d'enquête*, 14, 63-73.

SCOTT, A. J., et SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.

SCOTT, A. J., et HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SHAH, B. V. (1981a). SESUDAAN: Standard errors program for computing of standardized rates from sample survey data. Research Triangle Institute, Research Triangle Park, Caroline du Nord.

SHAH, B. V. (1981b). RATIOEST: Standard errors program for computing ratio estimates for sample survey data. Research Triangle Institute, Research Triangle Park, Caroline du Nord.

SHAH, B. V. (1982). RTIFREQS: Program to compute weighted frequencies, percentages and their standard errors. Research Triangle Institute, Research Triangle Park, Caroline du Nord.

SKINNER, C. J., HOLMES, D. J., et SMITH, T. M. F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.

SKINNER, C. J., HOLT, D., et SMITH, T. M. F. (1989). *Analysis of Complex Surveys*. New York, Wiley.

SMITH, T. M. F. (1984). Present position and potential developments: some personal views – sample surveys. *Journal of the Royal Statistical Society, series A*, 147, 208-221.

THOMAS, D. R., et RAO, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

TORTORA, R. D. (1980). The effect of disproportionate stratified design on principal component analysis used for variable elimination. *Proceedings of the Survey Research Section, American Statistical Association*, 746-750.

VERMA, V., et PEARCE, M. (1977). Users manual for CLUSTERS: A sampling program for computation of sampling errors for clustered samples. Technical Report No. 568, World Fertility Survey, U.K.

VINTER, S. (1980). Survey sampling errors with OSIRIS IV. COMPSTAT 1980: *Proceedings in Computational Statistics*, Vienna: Physica-Verlag, 72-80.

WANG, J. C., et WU, C. F. J. (1988). An approach to the construction of asymmetrical orthogonal arrays. Rapport technique, University of Waterloo.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York, Springer-Verlag.

WOODRUFF, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

WOODRUFF, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

WOODRUFF, R. S., et CAUSEY, B. D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

WU, C. F. J., HOLT, D., et HOLMES, D. J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, 3ième édition, London, Griffin.

- RAO, J.N.K. (1987). Analysis of categorical data from sample surveys, dans *New Perspectives in Theoretical and Applied Statistics* colligé par M.L. Puri, J.P. Vilaplana et W. Wertz), New York, Wiley, 45-60.
- RAO, J.N.K. (1988). Variance estimation in sample surveys, dans *Handbook of Statistics*, colligé par P.R. Krishnaiah et C.R. Rao), Vol. 6, Amsterdam, North-Holland, 427-447.
- RAO, J.N.K., et SINGH, M.P. (1973). On the choice of estimator in survey sampling. *Australian Journal of Statistics*, 15, 95-104.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K., et WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- RAO, J.N.K., et SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.
- RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, 231-241.
- RAO, J.N.K., et THOMAS D.R. (1988). The analysis of cross-classified categorical data from sample surveys. *Sociology Methodology*, 18, 213-269.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.
- ROYAL, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.
- ROYAL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- ROYAL, R.M., et HERSON, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 et 890-893.
- ROYAL, R.M., et EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, séries C, 37, 43-52.
- ROYAL, R.M., et PFEFFERMAN, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika*, 69, 401-410.
- ROYAL, R.M., et CUMBERLAND, W.G. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, series B, 50, 118-124.
- RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 4, 381-397.
- RYLETT, D.T., et BELLHOUSE, D.R. (1988). TREES: a computer program for complex surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 694-697.
- SÄRNDAL, C.E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- SÄRNDAL, C.E., et WRIGHT, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- SÄRNDAL, C.E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted regression technique for estimating the variance of the generalized regression estimator. *Biometrika*, 76, 527-537.
- SCHNELL, D., SULLIVAN, G., KENNEDY, W.J., et FULLER, W.A. (1986). PC CARP: Variance estimation for complex surveys, dans *Computer Science and Statistics: Proceedings of the 17th Symposium of the Interface*, colligé par D.M. Allen). Amsterdam: North Holland, 125-129.

- KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- LAVANGE, L.M., SHAH, B.V., BARNWELL, B.G., et KILLINGER, J.F. (1989). SUDAAN: A comprehensive package for survey data analysis. Rapport technique, Research Triangle Institute.
- LEPKOWSKI, J.M. (1982). The use of OSIRIS IV to analyse complex sample survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 38-43.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- MADOW, W.G. (1978). Comments on papers by Basu and Royall and Cumberland, dans *Survey Sampling and Measurement* colligé par N.K. Namboodiri. New York, Academic Press, 315-322.
- MADOW, W.G., et MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.
- MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-370.
- MCCARTHY, P.J. (1969). Pseudo-replication: half-samples. *International Statistical Review*, 37, 239-264.
- MOHADJER, T., MORGANSTEIN, D., CHU, A., et RHOADS, M. (1986). Estimation and analysis of survey data using SAS procedures WESVAR, NASSREG, and NASSLOG. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 258-263.
- MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.
- NARAIN, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.
- NATHAN, G. (1988). Inference based on data from complex sample designs, dans *Handbook of Statistics* colligé par P.R. Krishniah et C.R. Rao), Vol. 6, Amsterdam: North-Holland.
- NATHAN, G., et HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, séries B, 42, 377-386.
- NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, séries B, 12, 241-255.
- PFEFFERMAN, D., et SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.
- PLATT, W.G. (1986). GAUSS. *American Statistician*, 40, 164-169.
- RAJ, D. (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.
- RAO, J.N.K. (1971). Some thoughts on the foundations of survey sampling. *Journal of the Indian Society of Agricultural Statistics*, 23, 69-82.
- RAO, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators. *Journal of the Indian Statistical Association*, 17, 125-136.
- RAO, J.N.K. (1985). Inference conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11, 15-31.

- GODAMBE, V. P., et THOMPSON, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- GROSS, W. F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society*, séries B, 46, 270-272.
- GUPTA, V. K., et NIGAM, A. K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.
- HÄJKE, J. (1981). *Sampling From a Finite Population*. New York, Marcel Dekker.
- HANSEN, M. H., et HURWITZ, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- HANSEN, M. H., HURWITZ, W. N., et MADOW, W. G. (1953). *Sampling Survey Methods and Theory*, vol. 1. New York, Wiley.
- HANSEN, M. H., MADOW, W. G., et TEPPING, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- HANSEN, M. H., DALENUS, T., et TEPPING, B. J. (1985). The development of sample surveys of finite populations. In *A Celebration of Statistics: The ISI Centenary Volume* colligé par A. C. Atkinson and S. E. Fienberg), New York: Springer Verlag, 327-354.
- HARTLEY, H. O., et RAO, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- HIDIROGLOU, M. A., FULLER, W. A., et HICKMAN, R. (1980). *SUPERCARP-Sixth Edition*. Survey Section, Ames, Iowa.
- HIDIROGLOU, M. A., et PATON, D. J. (1987). Some experiences in computing estimates and their variances using data from complex survey designs, dans *Applied Probability, Statistics and Sampling Theory*, colligé par I. B. MacNeill et G. J. Umphrey), Boston, D. Reidel Publishing Company, 285-308.
- HOLT, D., et SMITH T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Séries A, 142, 33-46.
- HOLT, M. M. (1979). SURREG: standard errors of regression coefficients from sampling survey data. Research Triangle Institute, Research Triangle Park, Caroline du Nord.
- HORVITZ, D. G., et THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C. T., et FULLER, W. A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.
- JONES, G. K. (1983). HESBRR (HES variance and crossstabulation program). Version 3, Internal NCHS Report, Hyattsville, Maryland.
- KIAER, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.
- KISH, L. (1965). *Survey Sampling*. New York, Wiley.
- KISH, L., et FRANKEL, M. R. (1974). Inference from complex samples. *Journal of Royal Statistical Society*, series B, 36, 1-37.
- KOCH, G. G., FREEMAN, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KOVAR, J., RAO, J. N. K., et WU, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplément, 25-45.
- KOTT, P. S. (1987). Estimating the conditional variance of a design consistent regression estimator. Rapport technique.

- COCHRAN, W. G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.
- COCHRAN, W. G. (1946). Relative accuracy of systematic and stratified samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.
- COHEN, S. B., BURR, V. L., et JONES, G. K. (1986). Efficiencies in variance estimation for complex survey data. *American Statistician*, 40, 157-164.
- COHEN, S. B., XANTHOPOULIS, J. A., et JONES, G. K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data. *Journal of Official Statistics*, 4, 17-34.
- DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almqvist and Wiksell.
- DEMING, W. E. (1960). *Sample Design in Business Research*. New York, Wiley.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling, dans *New Developments in Survey Sampling*, colligé par N. L. Johnson et H. Smith, New York, Wiley-Interscience, 629-651.
- ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, series B*, 31, 195-224.
- FAY, R. E. (1982). Contingency tables for complex designs, CPLX. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 44-53.
- FAY, R. E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FELLEGI, I. P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.
- FIELDER, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428-440.
- FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd 5ième édition 1934.
- FOLSOM, R. E. (1974). National assessment approach to sampling error estimation, sampling error monograph. National Assessment of Educational Progress, première version.
- FRANCISCO, C. A., et FULLER, W. A. (1986). Estimation of the distribution function with a complex survey. Rapport technique, Iowa State University.
- FREEMAN, D. H. (1988). Sample survey analysis: analysis of variance and contingency tables, dans *Handbook of Statistics* colligé par P. R. Krishnaiah et C. R. Rao, vol. 6, Amsterdam: North-Holland, 415-426.
- FREEMAN, D. H., LIVINGSTON, M., LEO, L., et LEAF, P. (1985). A comparison of indirect variance estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 313-316.
- FULLER, W. A. (1975). Regression analysis for sample survey. *Sankhyā, Series C*, 37, 117-132.
- FULLER, W. A. (1987). Estimators of the factor model for survey data, dans *Applied Probability, Statistics and Sampling Theory* colligé par I. B. MacNeill et G. J. Umphrey, Boston: D. Reidel Publishing Company, 265-284.
- GHOSH, M. (1987). On admissibility and uniform admissibility in finite population sampling, dans *Applied Probability, Stochastic Processes and Sampling Theory*, colligé par I. B. MacNeill et G. J. Umphrey, Boston: D. Reidel Publishing Company, 265-284.
- GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, series B*, 17, 269-278.
- GODAMBE, V. P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, series B*, 28, 310-328.

mieux comprendre les plans de sondage complexes. Les progrès récents observés au chapitre de l'estimation de la variance et de la construction d'intervalles de confiance pour les statistiques non linéaires et les logiciels nouveaux créés à cette fin sont tout aussi impressionnants. Nous devons aussi nous réjouir des progrès rapides qui ont été faits au chapitre de l'analyse des données d'enquête; en effet, on a réussi à élaborer des méthodes d'analyse qui tiennent compte de la complexité du plan de sondage. Les progrès sont tout aussi réjouissants sur le plan des logiciels.

Dans les dix prochaines années, nous devrions assister à de nouvelles réalisations importantes en ce qui a trait à l'estimation de la variance pour les statistiques non linéaires (spécialement les fonctions non lisses), à l'analyse de données d'enquête (spécialement l'analyse multidimensionnelle) et à d'autres sujets qui n'ont pas été traités ici (notamment, l'échantillonnage dans le temps et l'estimation pour petites régions).

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur en chef pour ses commentaires utiles. Cette étude a pu être réalisée grâce à des subventions du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

- BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I, dans *Foundations of Statistical Inference*, colligé par V. P. Godambe et D. A. Sprott, Toronto, Holt, Rinehart et Winston, 203-242.
- BEBBINGTON, A. C., et SMITH, T. M. F. (1977). The effect of survey design on multivariate analysis, dans *The Analysis of Survey Data*, colligé par C. A. O'Muircheartaigh et C. D. Payne, vol. 2, New York, Wiley, 175-192.
- BEDRICK, E. J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.
- BELLHOUSE, D. R. (1988). A brief history of random sampling methods, dans *Handbook of Statistics*, colligé par P. R. Krishnaiah et C. R. Rao, vol. 6, Amsterdam: North-Holland, 1-14.
- BELLHOUSE, D. R., et RAO, J. N. K. (1986). On the efficiency of prediction estimators in two-stage sampling. *Journal of Statistical Planning and Inference*, 13, 269-281.
- BINDER, D. A. (1983). On the variance of the asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BOWLEY, A. L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6-62.
- BREWER, K. R. W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- BREWER, K. R. W., et HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.
- CASSEL, C. M., SÄRDAL, C. E., et WRETMAN, J. H. (1976). *Foundations of Inference in Survey Sampling*. New York, Wiley.
- CHAUDHURI, A. (1988). Optimality of sampling strategies, dans *Handbook of Statistics*, colligé par P. R. Krishnaiah et C. R. Rao, vol. 6, Amsterdam, North-Holland, 47-96.
- CHAUDHURI, A., et VOS, J. W. E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam, North-Holland.

définies par Francisco et Fuller (1986). En ce qui a trait aux données qualitatives, la version PC prévoit l'analyse de tableaux de contingence fondée sur les corrections de Rao-Scott appliquées au test chi carré d'indépendance. Le programme peut aussi servir à des analyses factuelles appliquées à des données d'enquête.

On compte quatre autres logiciels spécialisés pour l'analyse de données d'enquête; ils servent notamment à l'analyse de régression et à l'analyse de données qualitatives. Le programme &REPERR d'OSIRIS IV et la procédure SURREG de SUDAAN permettent de calculer l'erreur type de coefficients de régression, ce qui rend possible des analyses de régression. Les programmes CPLX (Fay 1982) et RSPLX, conçus tous deux par Fay, permettent de faire des analyses de données qualitatives de modèles linéaires logarithmiques pour des tableaux de contingence. Dans le premier cas, l'analyse se fait à l'aide du critère chi carré avec estimateur jackknife tandis que dans le deuxième cas, elle se fait au moyen du critère habituel, auquel ont été appliquées les corrections de second degré de Rao-Scott.

Les quatre programmes d'analyse de régression pour données d'enquête à plan de sondage complexe ont été évalués par Cohen, Xanthopoulos et Jones (1988). À cette occasion, on a choisi d'évaluer SUPER CARP plutôt que la version plus récente PC CARP. Comme dans l'étude de Cohen, Burt et Jones (1986) sur l'estimation de la variance, on s'est servi de données de la National Medical Care Expenditure Survey. L'évaluation a permis de constater que le logiciel qui exigeait le moins de temps processeur et qui était le plus facile à programmer était encore un logiciel de la série SUDAAN, notamment SURREG. Néanmoins, l'efficacité des programmes SUDAAN pourrait être contrebalancée par la souplesse du programme.

Le nouveau système SUDAAN qui est en voie d'élaboration (La Vange et coll. 1989) représente une amélioration notable par rapport à l'ancien système. Parmi les nombreuses modifications qu'il comporte, notons des méthodes d'estimation de la variance et d'analyse de données qui n'existaient pas auparavant dans l'ancien système.

On observe aussi une tendance, encore nouvelle, à intégrer des méthodes d'analyse de données d'enquête à plan de sondage complexe dans des logiciels et des systèmes statistiques courants. Après avoir calculé des estimations de la variance à l'aide de procédures SAS, Hidroglou et Paton (1987) décrivent d'autres procédures SAS qui, elles, permettent de faire des analyses log-linéaires (avec corrections de Rao-Scott) de tableaux de contingence. De même, Freeman (1988) signale qu'il a utilisé la procédure PROC MATRIX du SAS pour établir des estimations de variance et pour analyser la variance de ses données d'enquête. Enfin, Mohadjer et coll. (1986) décrivent deux nouvelles procédures SAS, outre WESVAR qui sert à l'estimation de la variance. Ce sont les procédures NASSREG et NASSLOG, qui servent à l'analyse de régression par les moindres carrés pondérés et à l'analyse de régression logistique respectivement. Dans les deux cas, l'estimation de la variance pour les paramètres de modèle se fait par la méthode BRR (balanced repeated replication). Notons la possibilité d'utiliser le langage d'algèbre matricielle GAUSS (Platt 1986) au lieu des procédures SAS. Se fondant sur leur propre expérience, Rao et Thomas (1988) vantent les mérites de ce langage pour l'analyse de données qualitatives dans les enquêtes à plan de sondage complexe.

6. CONCLUSIONS

Les premières tentatives pour élaborer des plans de sondage efficaces et les méthodes d'estimation de total et de moyenne de population qui s'y rattachent ont contribué à faire de la théorie des sondages une discipline importante de la statistique. Les progrès qui ont été réalisés par la suite dans cette théorie ont permis d'approfondir divers aspects de l'inférence statistique. En particulier, l'approche fondée sur un modèle et l'approche conditionnelle fondée sur un plan semblent offrir des perspectives intéressantes puisqu'elles visent à faire le lien entre l'approche classique et l'approche dépendante d'un modèle en intégrant les avantages de chacune; néanmoins, il faudra pousser plus loin la recherche dans ce domaine si l'on veut

(section 4.3). Fuller (1987) a calculé des estimateurs (fondés sur un plan) des paramètres de l'analyse factorielle ainsi que la matrice des covariances estimée correspondante. Il a montré que les variances estimées fondées sur la théorie normale peuvent être sensiblement inférieures aux variances réelles des estimateurs de facteurs.

5. LOGICIELS

Dans la seconde moitié des années 1970, on a élaboré de nombreux logiciels destinés à l'estimation de la variance dans les enquêtes complexes; en même temps, on développait des logiciels pour l'analyse de régression appliquée à des données d'enquête. Wolter (1985, p. 393-412) a passé en revue les programmes qui existaient vers 1985. Parmi ceux répertoriés par Wolter, les plus courants sont CLUSTERS (Verma et Pearce 1977), &PSALMS et &RPERR du système OSIRIS IV (Winter 1980 et Lepkowski 1982), SUDAN (Shah 1981a, 1981b, 1982 et Holt 1979), HESBRR (Jones 1983) et SUPER CARP (Hidiroglou, Fuller et Hickman 1980). Les programmes HESBRR et &RPERR utilisent la méthode BRR (balanced repeated replication) tandis que les autres utilisent la méthode de linéarisation de Taylor.

Cohen, Burt et Jones (1986) ont évalué les programmes d'estimation de la variance pour les moyennes et les quotients (à l'exception de CLUSTERS) au moyen d'une longue série de données de la National Medical Care Expenditure Survey. Ils ont constaté que les programmes SESUDAN et RATIOEST de la série SUDAN étaient ceux qui exigeaient le moins de temps processeur et qui étaient les plus faciles à programmer.

À l'heure actuelle, on conçoit de plus en plus de logiciels à base de menus pour micro-ordinateurs. Les programmes destinés à l'estimation de la variance et à l'analyse de données d'enquête n'échappent pas à cette tendance. D'ailleurs, l'introduction de PC CARP au milieu des années 1980 (Schnell et coll., 1986 et Schnell et coll., 1988) a nettement amélioré les programmes d'estimation de la variance qui étaient alors en usage; PC CARP peut être utilisé avec des micro-ordinateurs AT ou XT d'IBM ou des ordinateurs compatibles et exige un coprocesseur arithmétique. Comme son antécédent SUPER CARP, il utilise la méthode de linéarisation de Taylor pour l'estimation de la variance. Il existe un autre logiciel d'estimation de la variance pour micro-ordinateurs. Répertorié sous l'appellation BELLHOUSE dans l'ouvrage de Wolter (1985, p. 399), ce logiciel a été adapté pour les micro-ordinateurs IBM avec ou sans coprocesseur par Rylett et Bellhouse (1988) sous l'appellation TREES. Il sert à représenter par des structures arborescentes les plans de sondage à plusieurs degrés stratifié et permet d'établir des estimations de la variance à l'aide d'algorithmes de traversée, en mettant à profit des résultats généraux sur l'estimation de la variance dans l'échantillonnage à plusieurs degrés (voir section 3.1).

Une autre tendance dans l'information des méthodes d'estimation de la variance et des méthodes d'analyse de données d'enquête est l'intégration de logiciels d'enquêtes à de grands systèmes d'analyse statistique. Parmi les premiers systèmes du genre, qui datent du début des années 1980, notons le système SUDAN, composé d'une série de procédures SAS. Freeman et coll. (1985) et Hidiroglou et Paton (1987) ont utilisé la procédure PROC MATRIX de SAS pour calculer des estimations de la variance; les premiers ont eu recours à la méthode BRR et les seconds, à la méthode de linéarisation de Taylor. Mohadjer et coll. (1986) signalent la création d'une autre procédure SAS, WESVAR, pour calculer des estimations de la variance par la méthode BRR.

Il existe toute une série de logiciels et de techniques de calcul pour réaliser les analyses décrites dans la section 4. Parmi les logiciels spécialisés dont nous disposons, PC CARP est probablement le plus complet. Le programme original, SUPER CARP, avait pour but d'effectuer des analyses de régression qu'avait conçues Fuller (1975); la version PC permet de faire la même chose mais elle permet, en plus, des analyses de données qualitatives et des inférences sur la fonction de distribution cumulative et les quantiles correspondants selon des méthodes

de première espèce des tests courants. Wu, Holt et Holmes (1988) ont réalisé une étude systématique de l'incidence de l'échantillonnage à deux degrés sur la statistique F et ont proposé, en remplacement de la méthode itérative des moindres carrés généralisés (MCG), une correction pour le test F pour les cas où la corrélation intra-grappe est inconnue. Qu'il s'agisse de la méthode des MCG ou de la correction pour test F , il faut pouvoir identifier les grappes et cela n'est peut-être pas possible lorsque l'on se sert des données d'enquête pour l'analyse secondaire.

Si le modèle de régression comprend toutes les variables de plan z qui ont un rapport avec la variable dépendante (par ex. : variables indicatrices de strate et mesures de taille des unités) et que les erreurs ϵ_i sont indépendantes avec une variance σ^2 constante, l'analyse de régression est valable selon l'approche dépendante d'un modèle (Pfefferman et Smith 1985). Cependant, ce genre de modèles contiennent peut-être un trop grand nombre de paramètres pour pouvoir être utiles. De plus, les variables du plan peuvent ne pas représenter beaucoup d'intérêt pour l'utilisateur ou peuvent être inconnues dans l'analyse secondaire. Dans de tels cas, nous intéressons souvent à des modèles de la forme (4.1), où x n'est pas une variable de plan. En revanche, les paires d'échantillon (y_i, x_i) , $i \in s$, peuvent ne pas satisfaire le modèle à cause du biais d'échantillonnage. Nathan et Holt (1980) ont proposé une méthode de régression corrigée pour tenir compte du biais d'échantillonnage et ont comparé cette méthode à la méthode des moindres carrés ordinaires et à l'approche fondée sur un plan en utilisant B et $s(B)$ comme critères de comparaison. La méthode proposée suppose une relation particulière entre les variables de régression et les variables du plan. Les résultats empiriques de la comparaison indiquent que les inférences faites par la méthode des moindres carrés ordinaires peuvent être hautement incertaines, que l'approche fondée sur un plan est essentiellement fiable sauf dans des cas extrêmes, et que la méthode de régression corrigée est satisfaisante. Pfefferman et Holmes (1985) ont analysé la robustesse de ces méthodes, c'est-à-dire qu'ils ont vérifié jusqu'à quel point ces méthodes pouvaient être sensibles à une mauvaise définition des rapports entre les variables de régression, et ils en sont venus à la conclusion que la méthode de régression corrigée était très sensible. L'estimateur pondéré selon le plan B est robuste mais on peut obtenir un estimateur plus efficace en modifiant l'estimateur par régression corrigé de manière à en faire un estimateur convergent selon le plan pour le coefficient de régression de population finie B .

4.4 Analyse multidimensionnelle

Les méthodes mentionnées dans la section 4.2 pour l'analyse de moyennes de domaines peuvent s'appliquer aussi bien à des vecteurs de moyennes mais aucune étude détaillée sur la question ne figure dans les bibliographies statistiques. Les ouvrages qui traitent de l'analyse multidimensionnelle des données d'enquête ont surtout pour objet l'analyse des structures de covariance et plus particulièrement, l'analyse en composantes principales et l'analyse factorielle. Bebbington et Smith (1977), Tortora (1980) et Skinner, Holmes et Smith (1986) ont analysé l'effet du plan de sondage sur l'analyse en composantes principales. Les résultats de leur étude leur font conclure que le fait d'utiliser des méthodes standard sans prévoir un mécanisme de correction qui tienne compte du plan de sondage peut entraîner de fausses inférences. En particulier, les estimateurs des valeurs propres et des vecteurs propres de la matrice des covariances, Σ_y , peuvent être fortement biaisés dans le cas de plans de sondage non auto-pondérés. Skinner, Holmes et Smith (1986) ont proposé des estimateurs du maximum de vraisemblance (MV), selon un modèle normal multidimensionnel, ainsi que des estimateurs auto-pondérés. Les résultats de leur étude de simulation indiquent que les deux catégories d'estimateurs pondérés selon la probabilité sont entachés d'un biais de modèle conditionnel. En revanche, les estimateurs MV pondérés selon la probabilité peuvent toutefois être plus robustes, comme le montrent Pfefferman et Holmes (1985) en ce qui a trait à la méthode de régression corrigée

4.2 Analyse de moyennes ou de proportions de domaines

Les spécialistes des sciences sociales, des sciences de la santé et d'autres disciplines s'intéressent fortement à l'analyse de proportions de domaine (ou de sous-population) rattachées à une variable de réponse binaire. Dans le cas de proportions binomiales, on combine souvent les modèles de régression logistique avec les méthodes statistiques courantes. Rao et Scott (1987) ont déterminé les corrections du premier degré du test chi carré ordinaire pour la validité de l'ajustement et les hypothèses emboîtées; ces corrections peuvent être calculées à l'aide de tableaux publiés qui renferment des estimations des effets du plan (ou des erreurs types) des proportions de domaine. Roberts, Rao et Kumar (1987) ont déterminé les corrections du second degré (plus précises) des tests ordinaires mais rappelons-nous que pour exécuter ces tests, il faut connaître la matrice des covariances estimée des proportions de domaine. Ils ont élaboré du même coup des méthodes diagnostiques pour repérer les proportions aberrantes et les points déterminants dans l'espace-quotient tout en tenant compte du plan de sondage.

Koch, Freeman et Freeman (1975) ont eu recours à la méthode des moindres carrés pondérés pour analyser les moyennes de domaine d'une variable quantitative, y , et ont élaboré des tests de Wald pour vérifier la validité de l'ajustement du modèle et celle d'hypothèses linéaires portant sur les paramètres du modèle. Comme dans la section 4.1, il est possible d'accroître l'efficacité de ces tests en ayant recours à une modification impliquant la distribution F .

4.3 Analyse de régression linéaire

Dans la section 3.2, il a été question d'inférences fondées sur un plan qui portaient sur des paramètres de population finie non linéaires comme le coefficient de régression B d'une population finie. La variable-pivot $t = (B - B)/s(B)$ est distribuée approximativement selon une loi normale $N(0,1)$; B est l'estimateur convergent selon le plan (équation 3.2) de B et l'erreur type correspondante, $s(B)$ peut être calculée soit par la méthode de linéarisation (équation 3.4) ou par l'une ou l'autre des méthodes de ré-échantillonnage. Cette approche vaut aussi bien pour les coefficients de régression pondérée selon le plan B du coefficient de régressions courants permet de calculer l'estimateur pondéré selon le plan B du coefficient de régressions simple ou multiple; on utilise pour la circonstance l'inverse des poids de sondage qui se rattachent aux éléments de l'échantillon. Toutefois, l'erreur type de l'estimateur ainsi obtenu demeure inexacte.

Selon certains, la plupart des utilisateurs sont plus intéressés par les inférences portant sur les paramètres d'un modèle de superpopulation adéquat que par celles concernant des paramètres de population finie comme B . Néanmoins, B peut présenter de l'intérêt si on le définit comme l'estimateur par les moindres carrés du paramètre de superpopulation β dans le modèle

$$y_i = \alpha + \beta x_i + \epsilon_i \text{ with } E_m(\epsilon_i) = 0, \quad i = 1, \dots, N. \tag{4.1}$$

Si la taille de la population est élevée, estimer B revient effectivement à estimer β . En revanche, si le modèle (4.1) est mal défini de sorte que β perd toute signification, B présentera encore de l'intérêt comme pente de la droite des moindres carrés ajustée aux N -paires (y_i, x_i) (Godambe et Thompson 1986).

Scott et Holt (1982) ont adopté une approche dépendante d'un modèle pour étudier l'effet de l'échantillonnage à deux degrés sur l'analyse de régression. Comme dans Fuller (1975), ils ont supposé un modèle de régression de la forme (4.1) avec termes d'erreur ϵ_i équicorrélés dans chaque grappe. Ce modèle vaut aussi pour les paires d'échantillon (y_i, x_i) , i.e.s, si les probabilités de sélection n'ont aucun rapport avec la variable dépendante, comme c'est le cas pour l'échantillonnage aléatoire à deux degrés. Les résultats obtenus par Scott et Holt indiquent que l'existence d'une corrélation intra-grappe positive aura pour effet de sous-estimer l'erreur type des estimations de paramètres et d'exagérer par conséquent la probabilité d'erreur

(d) analyse multidimensionnelle, y compris l'analyse en composantes principales et l'analyse factorielle. Dans la présente section, nous faisons un bref compte rendu des progrès réalisés à chaque chapitre; à cet égard, nous invitons le lecteur à consulter les articles de synthèse de Nathan (1988), Rao (1987) et Smith (1984) de même que l'ouvrage de C.J. Skinner, D. Holt et T.M.F. Smith (1989).

4.1 Analyse de tableaux de contingence

Les tests chi carré (ou tests du rapport des vraisemblances) sont souvent utilisés pour l'évaluation et la sélection de modèles parcimonieux de \mathbf{p} , les probabilités par case de la population, dans un tableau de contingence de T cases. Les modèles linéaires logarithmiques sont tout à fait indiqués dans les circonstances parce que, comme dans l'analyse de variance, ils produisent constamment des critères pour tester diverses hypothèses se rapportant à un tableau à plusieurs dimensions. Rao et Scott (1984) ont réalisé une étude systématique des effets du plan de sondage sur le test chi carré de validité de l'ajustement d'un modèle log-linéaire, désigné par X^2 . Ils ont montré que X^2 est distribué asymptotiquement comme la somme pondérée, $\sum \delta_i W_i$, de $T - 1$ variables χ^2_1 indépendantes W_i , où les poids δ_i sont les valeurs propres d'une matrice des "effets du plan généralisés" et $T - 1$ représente le nombre de degrés de liberté. Ce résultat général montre que le plan de sondage peut avoir des effets appréciables sur la probabilité d'erreur de première espèce de X^2 . Par exemple, suivant un modèle d'échantillonnage de grappes produisant des effets de plan constants, $\delta_i = \lambda$ pour tous i , la probabilité d'erreur de première espèce réelle, pour un niveau nominal α , est approximativement $P_r[X^2_{T-r-1} > \lambda^{-1} \chi^2_{T-r-1}(\alpha)]$ et augmente avec l'effet de grappe, λ .

Rao et Scott (1984, 1987) ont obtenu des corrections du premier degré de X^2 ; ces corrections peuvent être établies à l'aide de tableaux publiés qui contiennent des estimations des effets du plan (ou des erreurs types) pour les estimations de case \mathbf{p} et les totaux marginaux correspondants, ce qui facilite les analyses secondaires faites à partir de tableaux publiés (voir aussi Gross, 1984 et Bedrick, 1983). Une correction du premier degré relie X^2/δ à χ^2_{T-r-1} où δ est la valeur estimée de l'effet du plan moyen $\delta = \sum \delta_i / (T - 1)$ ou la valeur estimée de la borne supérieure de δ . Le test corrigé est asymptotiquement valable pour un modèle produisant des effets de plan constants et devrait normalement donner des résultats satisfaisants lorsque la variabilité des δ_i 's est faible. On peut aussi obtenir des corrections du second degré, plus précises, qui tiennent compte de la variabilité des δ_i 's en appliquant la formule d'approximation de Satterthwaite à la somme pondérée de variables χ^2_1 indépendantes (Rao et Scott 1984). Toutefois, ces tests exigent que l'on connaisse la matrice complète des covariances estimées de \mathbf{p} . Parmi les autres méthodes qui tiennent compte du plan de sondage, notons le test de Wald fondé sur les moindres carrés pondérés (Koch, Freeman et Freeman 1975) et le test chi carré avec estimateur jackknife (Fay 1985). Ce dernier peut s'appliquer à des plans de sondage qui permettent l'utilisation de méthodes de ré-échantillonnage comme la méthode jackknife et la méthode BRR (balanced repeated replication). Le test de Wald exige la matrice complète des covariances estimées de \mathbf{p} , tandis que le test avec estimateur jackknife nécessite l'utilisation d'estimations de grappe ou de fichiers de micro-données.

Fay (1985) et Thomas et Rao (1987) ont montré que même si le test de Wald qui se rapporte à χ^2_{T-r-1} était asymptotiquement juste, il pouvait devenir très instable si le nombre de cases du tableau de contingence augmentait et que le nombre de grappes échantillonnées diminuait, cela ayant pour effet de porter à un niveau inacceptable (par rapport au niveau nominal α) la probabilité d'erreur de première espèce. Par ailleurs, le test de Fay et les corrections de Rao-Scott donnent des résultats satisfaisants dans des conditions très générales. Une version modifiée du test de Wald, qui se rapporte à une distribution F avec $T - r - 1$ et $T + r + 2$ degrés de liberté, s'est révélée plus efficace que le test de Wald en ce qui a trait à la bonne maîtrise de la probabilité d'erreur de première espèce (f représente le nombre de degrés de liberté requis pour estimer la matrice des covariances de \mathbf{p}).

détermine le niveau de confiance empirique des intervalles de confiance à $1 - \alpha$, $\theta \pm t_{\alpha/2}(\theta)$, pour les quotients et les coefficients de régression et de corrélation, où $t_{\alpha/2}$ est la limite supérieure ($\alpha/2$ pour cent) d'une variable t avec L degrés de liberté et $s^2(\theta)$ est l'un ou l'autre des estimateurs de la variance. Au point de vue du niveau de confiance empirique, la méthode BRR donne systématiquement de meilleurs résultats que la méthode jackknife, qui est elle-même supérieure à la méthode de linéarisation; les différences entre ces méthodes sont moins évidentes en ce qui a trait aux quotients. Au point de vue de la stabilité de l'estimateur de variance, on observe la situation inverse, c'est-à-dire que la méthode de linéarisation est la plus efficace, suivie de la méthode jackknife et de la méthode BRR. D'autres études empiriques rapportent des résultats similaires. Pour ce qui a trait à la méthode bootstrap, Kovar, Rao et Wu (1988) ont réalisé une étude de simulation qui indique que les intervalles bootstrap t reflètent mieux le taux d'erreur nominal à chaque extrémité que les intervalles fondés sur l'approximation normale de $t = (\theta - \theta)/s(\theta)$; en revanche, les estimateurs bootstrap sont moins stables que ceux fondés sur la méthode de linéarisation ou la méthode jackknife. Enfin, les études empiriques confirment que l'estimateur jackknife et l'estimateur de la méthode de linéarisation sont équivalents à des termes du second degré pour le cas particulier $n_h = 2$.

On trouve aussi dans les ouvrages de statistiques des méthodes d'estimation moins complexes que celles définies ci-dessus; notons par exemple la méthode des groupes aléatoires et la répétition partiellement équilibrée (variante de la méthode BRR). Cependant, les estimateurs de ces méthodes ne se ramènent pas à l'estimateur "standard" dans le cas de relations linéaires. Les ouvrages spécialisés proposent aussi des méthodes pour construire des modèles qui peuvent servir à imputer des erreurs d'échantillonnage. Ces méthodes sont un bon moyen d'obtenir des erreurs types "lissées" pour les estimateurs qui n'ont pas été calculés directement, et de représenter les erreurs types de façon concise (par des graphiques notamment) dans des rapports publiés.

L'ouvrage de Wolter (1985) est une excellente introduction à l'étude des méthodes d'estimation de la variance les plus récentes; ces méthodes y sont illustrées à l'aide de données provenant de diverses grandes enquêtes. Parmi les rapports de synthèses récents sur l'estimation de la variance, notons ceux de Rust (1985) et Rao (1988).

4. ANALYSE DE DONNÉES D'ENQUÊTE

Les méthodes courantes d'analyse de données reposent en règle générale sur l'hypothèse de l'échantillonnage aléatoire simple. Ces méthodes se retrouvent dans des logiciels statistiques courants comme SPSS^X, BMDP et SAS. Cependant, le fait d'appliquer des méthodes standard à des données d'enquête sans prévoir un mécanisme de redressement qui tienne compte du plan de sondage peut entraîner de fausses inférences puisque la plupart des données d'enquête sont recueillies au moyen de plans de sondage complexes qui comportent des échantillonnages en grappes, des stratifications et des échantillonnages avec probabilités inégales, et ne répondent donc pas à l'hypothèse de l'échantillonnage aléatoire simple. En particulier, on risque de sous-estimer fortement l'erreur type des estimations de paramètres et les intervalles de confiance correspondants si on ne tient pas compte de l'effet du plan de sondage dans l'analyse des données. De même, la probabilité d'erreur de première espèce des tests d'hypothèses peut être beaucoup plus élevée que la probabilité d'erreur nominale. Les types courants d'analyse de données préliminaires comme l'analyse des résidus, qui sert à déceler les écarts par rapport au modèle, n'échappent pas à cette condition. Kish et Frankel (1974) et d'autres auteurs ont mis en évidence quelques-uns des problèmes que posait l'utilisation des méthodes standard et ont fait valoir la nécessité de trouver de nouvelles méthodes qui tiennent vraiment compte de la complexité des plans de sondage. Au cours des dix dernières années, on n'a pas tardé à élaborer de telles méthodes pour les types d'analyse suivants: a) analyse de tableaux de contingence; b) analyse de moyennes ou de proportions de domaines; c) analyse de régression linéaire;

où $\hat{\theta}^{(j)}$ est l'estimateur calculé à l'aide du demi-échantillon j . Là encore, il existe plusieurs variantes de l'équation (3.6). On a étendu récemment la méthode BRR au cas général des n_h inégaux en utilisant des tableaux orthogonaux asymétriques (Gupta et Nigam 1987; Wang et Wu 1988). Pour l'échantillonnage stratifié, la méthode bootstrap comporte les étapes suivantes (Rao et Wu 1988): i) prélever avec remise, et de façon indépendante pour chaque h , un échantillon aléatoire simple $\{\mathbf{r}_{hi}^*\}_{i=1}^{m_h}$ de taille m_h dans $\{\mathbf{r}_h\}_{h=1}^H$. Calculer

$$\bar{\mathbf{r}}_h = \bar{\mathbf{r}}_h + [m_h/(n_h - 1)]^{1/2} (\mathbf{r}_{hi}^* - \bar{\mathbf{r}}_h), \bar{\mathbf{r}}_h = n_h^{-1} \sum_{i=1}^i \mathbf{r}_{hi}$$

et $\bar{\theta} = g(\Sigma \bar{\mathbf{r}}_h)$. ii) Répéter de façon indépendante l'étape i) un nombre élevé, B , de fois et calculer les estimateurs correspondants $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. iii) L'estimateur bootstrap de la variance de $\hat{\theta}$ est défini

$$s_{\hat{\theta}}^2 \text{BOOT}(\hat{\theta}) = \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta})^2 / (B - 1). \quad (3.7)$$

On peut aussi établir des intervalles de confiance en prenant comme approximation de la distribution de $t = (\hat{\theta} - \theta) / s_f(\hat{\theta})$ son équivalent bootstrap $\tilde{t} = (\hat{\theta} - \theta) / s_f^*(\hat{\theta})$, où $s_f^*(\hat{\theta})$ est déterminé à l'aide de $s_f^2(\hat{\theta})$ par l'application de la méthode jackknife à l'échantillon bootstrap particulier $\{\mathbf{r}_{hi}^*\}$. Ainsi, on peut définir pour θ les intervalles de confiance "bootstrap" à $1 - \alpha$ suivants

$$\{\hat{\theta} - t_{\text{UP}S_f}(\hat{\theta}), \hat{\theta} - t_{\text{LOW}S_f}(\hat{\theta})\}, \quad (3.8)$$

où t_{LOW} et t_{UP} sont les limites inférieure et supérieure (à $\alpha/2$ pour cent) de \tilde{t} tirées de l'histogramme bootstrap de t_1^*, \dots, t_B^* . Cet histogramme permet aussi de déterminer des intervalles de confiance unilatéraux. Par ailleurs, l'estimateur de la méthode de linéarisation peut remplacer l'estimateur jackknife pour la construction d'intervalles de confiance. Enfin, la construction d'intervalles de confiance exige un nombre beaucoup plus élevé, B , d'échantillons bootstrap que l'estimation de la variance. En ce qui concerne la taille de l'échantillon bootstrap m_h la plus appropriée, $m_h = n_h - 1$ est un choix intéressant puisqu'on obtient $\bar{\mathbf{r}}_{hi}^* = \bar{\mathbf{r}}_h$.

Comparaison des méthodes

Les méthodes exposées dans les ouvrages de statistique ont les propriétés théoriques suivantes: 1) Tous les estimateurs de variance se réduisent à l'estimateur "standard" $s^2(\hat{Y})$, défini en (3.1), dans le cas de la relation linéaire $g(\mathbf{Y}) = Y$. 2) Pour les fonctions lisses $g(\mathbf{Y})$, tous les estimateurs de variance sont asymptotiquement convergents selon le plan (Krewski et Rao 1981). Toutefois, l'estimateur jackknife est réputé non convergent pour des fonctions non lisses comme les quantiles, même lorsqu'il s'agit d'un échantillonnage aléatoire simple. Par conséquent, il conviendra d'utiliser les logiciels "jackknife" avec prudence. 3) Si $n_h = 2$ pour tous h l'estimateur jackknife et l'estimateur de la méthode de linéarisation sont asymptotiquement équivalents aux termes de degré supérieur pour les fonctions lisses $g(\mathbf{Y})$, ce qui indique que le choix de l'une ou l'autre méthode dans ce cas très particulier devrait dépendre plus de facteurs tels le coût des calculs (Rao et Wu 1985). Pour ce qui a trait aux études empiriques, Kish et Frankel (1974) ont analysé la méthode de linéarisation de même que les méthodes jackknife et BRR à l'aide de données de la Current Population Survey et de plans de sondage prévoyant le tirage de $n_h = 2$ grappes dans chacune des strates $L = 6, 12$ et 30 . Ils ont

tandis que la méthode bootstrap semble avoir une application plus générale. En revanche, le bootstrap est plus complexe sur le plan du calcul et ses propriétés n'ont pas encore été entièrement explorées.

Méthode de linéarisation

Si nous désignons par $v(z_i)$ l'estimateur de la variance de $v(z_i)$ pour un plan général, la méthode de linéarisation produira un estimateur de la variance $v(z_i)$ pour une statistique non linéaire θ , z_i étant une variable synthétique définie convenablement, qui dépend de la forme de θ . Pour une statistique générale $\theta = g(\mathbf{Y})$, l'estimateur de la variance est défini

(3.3)
$$s^2_L(\theta) = v(z_i) \text{ with } z_i = \sum^i y_{ii} g_i(\mathbf{Y}),$$

(Woodruff 1971), où y_{ii} est la valeur de la caractéristique i pour l'unité i , et $g_i(\mathbf{Y})$ est la dérivée partielle $\partial g(\mathbf{Y}) / \partial Y_i$ évaluée à $\mathbf{Y} = \mathbf{Y}(i = 1, \dots, q)$. Un inconvénient de la formule (3.3) est que les dérivées partielles peuvent parfois être difficiles à évaluer, quoiqu'il est possible d'obtenir de bonnes approximations de ces dérivées à l'aide de méthodes numériques (Woodruff et Causey 1976). Il arrive souvent que l'on puisse obtenir l'estimateur de la variance sans devoir calculer les dérivées partielles g_i ; il suffit pour cela de redéfinir la statistique θ comme un quotient et d'utiliser la formule de variance qui s'applique habituellement aux quotients. Par exemple, on peut exprimer le coefficient de régression d'échantillon B par la formule $B = Y(z_{1i}) / Y(z_{2i})$ où $1_i = (y_i - \bar{y}) (x_i - \bar{x})$ et $2_i = (x_i - \bar{x})^2$, de telle sorte que

(3.4)
$$s^2_L(B) = v(z_{1i} - B z_{2i}) / [Y(z_{2i})]^2.$$

On peut se servir des mêmes méthodes pour d'autres statistiques comme les coefficients de régression multiple (Fuller 1975; Folsom 1974). Binder (1983) a étendu la méthode de linéarisation aux statistiques qui sont définies implicitement comme la solution d'un système d'équations non linéaires. Son exposé porte sur les paramètres de population finie tirés des modèles linéaires généralisés, qui comprennent le modèle de régression linéaire et le modèle de régression logistique.

Méthodes de ré-échantillonnage

Nous allons maintenant étudier l'application de méthodes de ré-échantillonnage dans le cas d'un plan de sondage à plusieurs degrés stratifié comme celui de la section 3.1. Si nous définissons θ^{hi} comme l'estimateur de θ calculé à l'aide de l'échantillon $\{\mathbf{r}_{hi}\}$, abstraction faite de $\mathbf{r}_{hi} = \mathbf{Y}_{hi}/p_{hi}$, l'estimateur jackknife de la variance de $\theta = g(\sum \mathbf{r}_{hi})$ est défini

(3.5)
$$s^2_J(\theta) = \sum^h \{ (n_h - 1) / n_h \} \sum^i (\theta^{hi} - \theta)^2.$$

Il existe plusieurs variantes de l'équation (3.5); par exemple, θ peut être remplacé par $\theta^h = \sum^i \theta^{hi} / n_h$.

McCarthy (1969) a proposé d'appliquer la méthode BRR pour le cas particulier où $n_h = 2$. On constitue un ensemble de J demi-échantillons "équilibrés" en supprimant une u.p.c. dans chaque strate de l'échantillon. Cet ensemble peut être formé à l'aide des matrices de Hadamard. L'estimateur BRR de la variance est défini

(3.6)
$$s^2_{\text{BRR}}(\theta) = \sum^J (\theta^{(j)} - \theta)^2 / J,$$

qui appartiennent à la catégorie générale d'estimateurs linéaires de Godambe, $Y_b = \sum_{i \in b} b_{is} Y_i$; cette approche simplifie le calcul de l'erreur quadratique moyenne et expose la forme fondamentale de n importe quel estimateur non biaisé quadratique non négatif de l'erreur quadratique moyenne. Pour les plans à plusieurs degrés, un estimateur général de Y est défini par une équation de la forme $Y_{bm} = \sum_{i \in b} b_{is} Y_i$, où s désigne maintenant un échantillon d'unités primaires d'échantillonnage (u.p.é.) et Y_i est un estimateur linéaire non biaisé du total d'u.p.é. Y_i établi en fonction d'un sous-échantillonnage de l'u.p.é. Raj (1966) et Rao (1975) ont élaboré des formules de variance uniformes pour les plans à plusieurs degrés.

Dans les grandes enquêtes, on utilise souvent un grand nombre de strates, L , mais un nombre relativement petit d'u.p.é., n_h , dans chaque strate h . De fait, il est d'usage de choisir $n_h = 2$ u.p.é. dans chaque strate de manière à réaliser une stratification maximum des u.p.é., tirées (avec remise) de la strate h avec une probabilité $n_h = 2$, l'estimateur du total Y est $Y = \sum_h Y_h$, et on obtient un estimateur non biaisé de la variance de cet estimateur par la formule

$$s^2(Y) = \sum_h \left\{ \sum_i (r_{hi} - f_h)^2 / [n_h(n_h - 1)] \right\}, \quad (3.1)$$

où $f_h = \sum_i r_{hi} / n_h$, $r_{hi} = Y_{hi} / p_{hi}$ et Y_{hi} est un estimateur non biaisé du total de l'u.p.é. i dans la strate h ($i = 1, \dots, n_h$; $h = 1, \dots, L$). On utilise souvent ce plan stratifié pour comparer des méthodes qui s'appliquent à des statistiques non linéaires (section 3.2). La simplicité de $s^2(Y)$ fait qu'il est souvent utilisé, même lorsque les u.p.é. sont tirées sans remise. Dans un tel cas, il y a surestimation de la variance mais le biais relatif sera peu élevé si la fraction de sondage du premier degré est faible.

3.2 Statistiques non linéaires

De nombreux paramètres de population finie non linéaires θ comme les rapports, les coefficients de régression et les coefficients de corrélation, peuvent être exprimés comme des fonctions lisses, $g(Y)$, de totaux $Y = (Y_1, \dots, Y_q)'$ de variables aléatoires convenablement définies, telles que $g(Y) \propto g_1(Y_1/M, \dots, Y_{q-1}/M)$ ou $Y^q = M$, la taille de la population. Le paramètre θ est estimé par $g(\hat{Y}) \propto g_1(\hat{Y}_1/M, \dots, \hat{Y}_{q-1}/M)$. Ces estimateurs sont connues même lorsque les variables aléatoires rattachées aux éléments i n'ont aucun rapport avec les probabilités de sélection π_i ($i = 1, \dots, N$) puisque $g(\hat{Y})$ n'est une fonction que des estimateurs de Hajek $\hat{Y}_j = \hat{Y}_j/M$ des moyennes \hat{Y}_j . Par exemple, on peut exprimer l'estimateur du coefficient de régression d'une population finie $B = \sum (x_i - \bar{X})(y_i - \bar{Y}) / \sum (x_i - \bar{X})^2$ par la formule

$$B = [Z/M - (X/M)(Y/M)] / [W/M - (X/M)^2]^{-1}, \quad (3.2)$$

où X , Z et W sont les estimateurs des totaux X , Z et W des x_i , $z_i = y_i x_i$ et $w_i = x_i^2$ respectivement.

Les méthodes d'estimation de la variance pour les statistiques non linéaires, $g(\hat{Y})$, comprennent la fameuse méthode de linéarisation ainsi que les méthodes de ré-échantillonnage comme le jackknife, la méthode BRR (balanced repeated replication) et le bootstrap. La méthode de linéarisation peut s'appliquer à des plans de sondage généraux mais elle exige une formule de variance distincte pour chaque statistique. En revanche, les méthodes de ré-échantillonnage utilisent la même formule pour toutes les statistiques. Par ailleurs, les méthodes jackknife et BRR s'appliquent exclusivement aux plans de sondage qui prévoient un échantillonnage avec remise des u.p.é. (ou des fractions de sondage du premier degré négligibles).

est d'ajouter au modèle (2.4) une variable auxiliaire $u_i = \sigma_i^2(1 - \pi_i)/\pi_i$, c'est-à-dire d'utiliser l'équation $E(y_i) = \beta x_i + \gamma u_i$ (Særdal et Wright 1984). Si nous ajoutons au même modèle deux variables auxiliaires, σ_i^2/π_i et σ_i^2 (ce qui donne $E(y_i) = \beta x_i + \gamma \sigma_i^2/\pi_i + \delta \sigma_i^2$, nous obtenons un meilleur estimateur linéaire non biaisé selon le modèle, asymptotiquement convergent selon le plan de la forme $\bar{Y} = \sum_{i \in S} g_{Si} y_i / \pi_i$ (Særdal et Wright 1984). De plus, nous observons une variance asymptotique espérée minimum si nous choisissons un plan d'échantillonnage où π_i est proportionnel à σ_i . En revanche, ces conditions idéales ne peuvent être réalisées sans une légère augmentation de la variance de modèle selon le modèle original (2.4).

Godambe et Thompson (1986) ont eu recours à la théorie des fonctions d'estimation pour déterminer des estimateurs convergents selon le plan à l'aide d'un modèle hypothétique. Par exemple, si, pour une caractéristique quelconque y étudiée dans une enquête polyvalente, il n'est pas censé exister de rapport entre y_i et π_i la fonction d'estimation "optimale" donne l'estimateur de Hájek (1971) de \bar{Y} :

$$\hat{\bar{Y}}_H = \left(\sum_{i \in S} y_i / \pi_i \right) / \left(\sum_{i \in S} 1 / \pi_i \right). \quad (2.11)$$

Le modèle de superpopulation est défini en l'occurrence par l'équation $y_i = \theta + \epsilon_i$, étant des erreurs indépendantes, ce qui reflète bien le cas qui nous occupe. L'estimateur \bar{Y}^H ne présente pas les inconvénients de l'estimateur d'Horvitz-Thompson $\bar{Y}^{HT/N}$, comme le montre Basu (1971) dans son exemple des "éléphants". La méthode des fonctions d'estimation offre des perspectives intéressantes mais il faut poursuivre les recherches sur la manière de l'utiliser pour obtenir de "meilleurs" estimateurs ou de "meilleures" variables-pivots ou les deux à la fois. Chose intéressante, cette méthode équivaut essentiellement à la fameuse méthode de Fieller qui sert à calculer des limites de confiance pour les quotients (Fieller 1932), et à la méthode de Woodruff (1952), par laquelle on calcule des limites de confiance pour les médianes.

Dans les sections 2.2 et 2.3, nous avons utilisé des modèles qui s'appliquent à l'échantillonnage à un degré. En ce qui concerne l'échantillonnage à plusieurs degrés, les modèles sont plus complexes à cause des corrélations intra-grappe (Scott et Smith 1969; Montanari 1987). Les meilleurs estimateurs (ou prédicteurs) linéaires non biaisés selon le modèle que l'on obtient dans ce cas sont des combinaisons pondérées d'estimateurs, où les poids dépendent des corrélations intra-grappe, celles-ci pouvant être estimées à l'aide des données de l'échantillon. Bellhouse et Rao (1986) ont étudié l'efficacité relative de ces estimateurs dans le cas d'un échantillonnage répété. Les résultats empiriques de leur étude donnent à penser que les prédicteurs pourraient ne pas être beaucoup plus efficaces que l'estimateur habituel dans un échantillonnage à deux degrés où il y a d'abord échantillonnage de grappes avec PPT, puis échantillonnage aléatoire simple à l'intérieur des grappes échantillonnées.

Si les grappes sont considérées comme des strates et que les moyennes de strates sont les paramètres d'intérêt comme dans l'estimation pour petites régions, les prédicteurs de moyennes de strates devraient être beaucoup plus efficaces que les estimateurs fondés sur un plan habituels puisqu'ils, contrairement à ceux-ci, ils "tirent parti" de toutes les strates. S'il s'agit d'un échantillonnage à deux degrés où les moyennes de grappes sont les paramètres d'intérêt, il ne peut être question que d'un prédicteur pour les grappes non échantillonnées.

3. ESTIMATION DE LA VARIANCE ET INTERVALLES DE CONFIANCE

3.1 Statistiques linéaires

Une bonne partie de la théorie classique des sondages est consacrée au calcul de l'erreur quadratique moyenne ou de la variance d'estimateurs linéaires d'un total Y et au calcul de l'estimateur de cette variance. Rao (1979) a élaboré une approche uniforme pour les estimateurs

$$Y^{reg} = \sum_{i \in S} y_i / \pi_i + \beta \left(X - \sum_{i \in S} x_i / \pi_i \right) \quad (2.7)$$

est asymptotiquement optimal (c'est-à-dire que la variance espérée asymptotique atteint la borne inférieure) pour n importe quel plan à taille d'échantillon fixe où σ_i est proportionnel à π_i . Dans l'équation ci-dessus, β est un estimateur linéaire non biaisé selon le modèle de β et $E_p^m(\beta - \beta)^2 \rightarrow 0$ lorsque $n \rightarrow \infty$ (Särndal 1980). Plus particulièrement, on peut choisir le meilleur estimateur non biaisé selon le modèle $\beta = (\sum_{i \in S} w_i x_i y_i) / (\sum_{i \in S} w_i x_i^2)$ où $w_i = 1/\sigma_i^2$.

Si on choisit $\hat{\beta} = (\sum_{i \in S} w_i x_i y_i / \pi_i) / (\sum_{i \in S} w_i x_i^2 / \pi_i)$, où $w_i = 1/x_i$, Y^{reg} se réduit à l'expression plus simple (estimateur par quotient)

$$Y^{reg} = X \hat{\beta} = \sum_{i \in S} g_{si} y_i / \pi_i, \quad (2.8)$$

où $g_{si} = X / (\sum_{i \in S} x_i / \pi_i)$ et g_{si} et converge en probabilité vers 1 lorsque $n \rightarrow \infty$ (Särndal et Wright 1984). Särndal, Swensson et Wretman (1989) ont proposé, pour les estimateurs Y du type (2.8), un nouvel estimateur de variance qui est à la fois convergent selon le plan et approximativement non biaisé pour la variance conditionnelle $V_m(Y - X)$. L'estimateur de variance qu'ils proposent pour Y^{reg} est défini

$$s^2(Y^{reg}) = \sum_{i < j \in S} (\pi_i \pi_j - \pi_{ij}) \pi_{ij}^{-1} (g_{si} \hat{e}_i - g_{sj} \hat{e}_j)^2 \quad (2.9)$$

où $\hat{e}_i = (y_i - \beta x_i) / \pi_i$. Dans le cas de l'échantillonnage aléatoire simple, $s^2(Y^{reg})$ se ramène à $s_a^2(Y)$, défini en (3.7), ce qui est justifié selon l'approche de la prédiction et l'approche de la randomisation conditionnelle. Kott (1987) a proposé d'appliquer un facteur de redressement à l'estimateur de la variance de Yates-Grundy classique, $s_{YG}^2(Y)$ de n importe quel estimateur Y fondé sur un modèle et asymptotiquement convergent selon le plan. L'estimateur de la variance de Kott

$$s_{YG}^2(Y) = s_{YG}^2(Y) [V_m(Y) - Y] / E_m s_{YG}^2(Y) \quad (2.10)$$

est à la fois non biaisé selon le modèle et asymptotiquement convergent selon le plan. Toutefois, pour des estimateurs du type (2.8), l'estimateur de la variance de Särndal et coll. semble plus simple puisqu'on l'obtient en remplaçant simplement \hat{e}_i par $g_{si} \hat{e}_i$ dans la formule de l'estimateur de variance classique $s_{YG}^2(Y)$.

On obtient l'estimateur par régression classique en considérant tout d'abord une constante B au lieu de β dans l'équation (2.7), puis en substituant à cet estimateur un estimateur convergent de B_{opt} , qui est la valeur de B pour laquelle la variance du plan est minimisée. L'estimateur par régression classique ne dépend pas de la validité d'aucun modèle. Cependant, on peut obtenir une valeur approchée de la variance de plan optimale dans l'approche fondée sur un modèle en redéfinissant le modèle (2.4) par l'équation $E(y_i) = \beta x_i + \gamma \pi_i$, puis en se servant de $(\hat{\beta}, \hat{\gamma})$, l'estimateur par régression pondérée de (β, γ) , avec comme poids $w_i = 1/\pi_i^2$. L'estimateur de Y ainsi obtenu se ramène à l'équation (2.7), où $\hat{\beta}$ est remplacé par $\hat{\beta}$ (Isaki et Fuller 1982; Montanari 1987). Toute autre valeur pour $\hat{\beta}$ donnera une variance asymptotique plus élevée.

Little (1983) a soutenu qu'il ne fallait utiliser que des modèles qui produisent des meilleurs estimateurs linéaires non biaisés selon le modèle et asymptotiquement convergents selon le plan puisque ces estimateurs sont optimaux lorsque le modèle est vrai. Une façon d'en arriver à cela

enquêtes où il y a plusieurs caractéristiques d'intérêt puisqu'on peut avoir besoin d'un échantillon différent pour chaque variable.

Si les variables concomitantes supplémentaires z du modèle ne sont pas connues ou ne sont pas mesurées, Royall et Pfeffermann (1982) recommandent l'échantillon choisi n'est pas mal équilibré par rapport à z . Toutefois, dans un article plus récent, Royall et Cumberland (1988) semblent être favorables à une forme de randomisation restreinte, comme le montre l'extrait suivant: "De nombreuses méthodes, parmi lesquelles la randomisation restreinte, la stratification et l'échantillonnage systématique, peuvent servir à obtenir des échantillons équilibrés. Nous n'avons de préférence pour aucune d'entre elles (. . .)". Quoi qu'il en soit, la plupart des partisans de l'approche dépendante d'un modèle semblent recommander l'échantillonnage probabiliste sous une forme ou sous une autre, comme le fait remarquer Smith (1984). Par conséquent, l'approche de l'échantillonnage probabiliste et l'approche dépendante d'un modèle se distinguent principalement par le choix de la variable-pivot qui contient l'estimateur Y ainsi qu'une mesure de sa variance.

Malgré les limites que nous venons d'exposer, l'approche dépendante d'un modèle est utile pour analyser l'efficacité conditionnelle des méthodes classiques suivant divers modèles plan-sibles. Par exemple, l'estimateur de la variance $s^2_r(Y_r)$ a le même comportement que la variance conditionnelle $V_m(Y_r - Y)$ suivant le modèle (2.4), où $\sigma^2_r = \sigma^2_{x_i}$, tandis que $s^2_r(Y_r)$ est biaisé selon le modèle (Royall et Eberhardt 1975). L'estimateur $s^2_r(Y_r)$ montre aussi de la robustesse lorsqu'on s'écarte de l'hypothèse $\sigma^2_r = \sigma^2_{x_i}$.

2.3 Approche fondée sur un modèle

Hansen, Madow et Tepping (1983) ont fait ressortir les inconvénients que comporte l'utilisation de méthodes dépendantes selon le plan qui sont aussi non biaisés selon le modèle conformément à un modèle hypothétique. On construit aussi des estimateurs de variance qui sont à la fois convergents, en ce qui a trait à la variance de plan de Y , et non biaisés selon le modèle (du moins approximativement), en ce qui a trait à la variance conditionnelle $V_m(Y - Y)$. Dans les circonstances, la variable-pivot permet donc de faire des inférences valables selon un modèle hypothétique et de parer aux défauts de sensibilité du modèle, notamment en favorisant des inférences s'est très peu intéressé aux propriétés conditionnelles fondées sur un plan des méthodes fondées sur un modèle dans l'hypothèse de l'existence de défauts de sensibilité.

Dans l'approche fondée sur un modèle, on ne s'intéresse qu'aux estimateurs Y asymptotiquement convergents selon le plan qui sont aussi non biaisés selon le modèle conformément à un modèle hypothétique. On construit aussi des estimateurs de variance qui sont à la fois convergents, en ce qui a trait à la variance de plan de Y , et non biaisés selon le modèle (du moins approximativement), en ce qui a trait à la variance conditionnelle $V_m(Y - Y)$. Dans les circonstances, la variable-pivot permet donc de faire des inférences valables selon un modèle hypothétique et de parer aux défauts de sensibilité du modèle, notamment en favorisant des inférences s'est très peu intéressé aux propriétés conditionnelles fondées sur un plan des méthodes fondées sur un modèle dans l'hypothèse de l'existence de défauts de sensibilité.

Godambe (1955) a posé par hypothèse le modèle (2.4), avec $V_m(Y_i) = \sigma^2_i$ et $\text{cov}_m(Y_i, Y_j) = 0, i \neq j$, et a déterminé une borne inférieure, $\sum^{i \in U} (1/\pi_i - 1)\sigma^2_i$, pour la variance prévue de n importe quel estimateur linéaire non biaisé selon le plan, Y_b . Il a de plus montré qu'en combinant un plan avec taille d'échantillon fixe où $\pi_i = (nx_i)/X$ avec l'estimateur d'Horvitz-Thompson, $Y_{HT} = \sum^{i \in S} y_i/\pi_i$, on atteint la borne inférieure à la condition que $\sigma^2_i = \sigma^2_{x_i}$. Si $\sigma^2_i \neq \sigma^2_{x_i}$, il n'existe pas de méthode "optimale" non biaisée selon le plan. C'est pourquoi on a élaboré des méthodes asymptotiquement optimales en considérant désormais, outre les estimateurs non biaisés selon le plan, les estimateurs asymptotiquement convergents selon le plan. L'estimateur par régression généralisé

L'approche dépendante d'un modèle a été mise de l'avant par Brewer (1963), puis approfondie par Royall et ses collaborateurs à partir de 1970 (voir Royall 1970). La meilleure façon de la représenter est par un modèle de régression simple

(2.4) $E_m(y_i) = \beta x_i, \quad i = 1, \dots, N; \quad \beta > 0, x_i > 0$

où E_m désigne l'espérance mathématique du modèle. On suppose en outre que la variance du modèle $V_m(y_i) = \sigma_i^2$ où σ_i^2 est connue sauf pour une constante multiplicative, et que la covariance du modèle $\text{cov}_m(y_i, y_j) = 0, i \neq j$. Royall (1970) a montré que l'estimateur ordinaire $N\bar{y}$, non biaisé selon un échantillonnage aléatoire simple, est biaisé selon le modèle défini en (2.4) et qu'il produit une forte sous-estimation si l'échantillon observé renferme surtout des unités de petite taille, x_i . On peut observer les mêmes résultats avec l'approche conditionnelle fondée sur un plan sans supposer de modèle (Rao 1985).
Le meilleur estimateur (ou prédicteur) linéaire non biaisé selon le modèle de Y suivant le modèle (2.4) est défini

(2.5)
$$Y = \sum_{i \in S} y_i + \sum_{i \notin S} \hat{\beta} x_i$$

où $\bar{s} = U - s$ est l'ensemble des unités non échantillonnées et $\hat{\beta}$ le meilleur estimateur linéaire sans biais de β . L'équation (2.5) se ramène à la formule de l'estimateur par quotient ordinaire \bar{Y}_r si $\sigma_i^2 = \sigma^2 x_i$. On calcule l'incertitude de Y au moyen de l'équation $E_m(Y - Y)^2 = V_m(Y - Y)$, qui devient

(2.6) $V_m(Y - Y) = \{X(X - nx)/(nx)\} \sigma^2$

lorsqu'il s'agit de l'estimateur par quotient ordinaire (Y_r). Comme (2.6) diminue lorsque x augmente, le plan optimal est un échantillonnage par choix raisonné des n unités ayant les valeurs x les plus élevées (en supposant que les valeurs x_i de la population sont connues). On peut obtenir un estimateur non biaisé selon le modèle, $s_y^2(Y - Y)$, de $V_m(Y - Y)$ à l'aide de l'équation (2.6) en remplaçant σ^2 par l'estimateur par les moindres carrés pondérés $\hat{\sigma}^2$, on obtient ainsi une variable-pivot $t_m = (Y - Y)/s_m(Y - Y)$ qui est distribuée approximativement selon une loi $N(0,1)$ d'après la distribution du modèle. Ces résultats théoriques sont impressionnants mais ce genre d'approche peut engendrer des biais appréciables si le modèle en question n'est pas entièrement correct.

Pour parer à cette difficulté, Royall et Herson (1973) ont considéré des écarts au modèle qui consistent en des termes polynomiaux du second degré ou d'un degré plus élevé en x (disons q -ième degré) ou en une ordonnée à l'origine ou les deux à la fois, et ont montré qu'un échantillon équilibré pour lequel $x^{(j)} = X^{(j)}, j = 1, \dots, q$, favorise la robustesse d'estimateurs en ce sens que \bar{Y}_r demeure non biaisé selon le modèle, $x^{(j)} = \sum_{i \in U} x_i^{(j)}/n$ et $X^{(j)} = \sum_{i \in U} x_i^{(j)}/N$. Ils ont aussi montré que l'on pouvait accroître l'efficacité de l'estimation en combinant l'utilisation de l'estimateur par quotient de Y avec une stratification en fonction de x avec répartition optimale et échantillonnage équilibré dans chaque strate. Néanmoins, les échantillons équilibrés qui sont choisis de façon raisonnée ont leurs inconvénients. Premièrement, l'auto-évaluation de l'échantillonnage équilibré dans chaque strate. Deuxièmement, l'échantillonnage est une opération qui ne présente plus les mêmes caractéristiques lorsqu'on s'écarte du modèle de régression polynomial (Madow 1978, p.320). Le modèle alternatif doit faire l'objet d'un équilibrage et il se peut que ce modèle renferme des termes polynomiaux de degré supérieur ou d'autres variables ou les deux à la fois et ces variables supplémentaires doivent être connues d'avance. Troisièmement, l'échantillonnage équilibré n'est pas réalisable dans les

quel plan d'échantillonnage; cependant, le bien-fondé des propriétés d'optimalité définies par Godambe soulève des interrogations (voir Rao 1971; Rao et Singh 1973). Par son fameux exemple des "éléphants", Basu (1971) prouve l'inutilité de deux critères en particulier, soit celui de la variance minimum "indispensable" et celui de l'hypermisibilité.

Godambe (1966) a déterminé la fonction de vraisemblance du paramètre d'intérêt $y = (y_1, \dots, y_N)'$ à partir de l'échantillon $\{(y_i, x_i), i \in s\}$ mais on ignore tout des valeurs (y_i, x_i) pour $i \notin s$, et du total X parce que les N unités de la population sont considérées essentiellement comme N strates d'échantillon indépendantes. Pour remédier à ce problème, on peut faire abstraction de quelques-unes des données de manière que l'échantillon ne soit plus unique et que l'on obtienne une fonction de vraisemblance informative (Hartley et Rao 1968; Royall 1968). Une autre façon de contourner la difficulté est de combiner la fonction de vraisemblance non informative avec des distributions a priori interchangeables à l'aide du théorème de Bayes pour en arriver à des inférences a posteriori informatives (Ericson 1969).

L'inférence conditionnelle est un sujet qui a reçu beaucoup d'attention (et fait l'objet d'une grande controverse) en statistique classique depuis Fisher (1925). Le choix d'un ensemble fondamental pertinent pour l'inférence conditionnelle ne s'impose pas toujours à l'évidence mais dans le cas de la stratification a posteriori, il semble raisonnable d'utiliser la taille effective des strates comme argument de condition dans l'inférence fondée sur un plan (Durbin 1969). Holt et Smith (1979) se font les propagandistes les plus ardents de l'inférence conditionnelle fondée sur un plan quoique leur analyse se limite à la stratification a posteriori d'un échantillon aléatoire simple. Rao (1985) examine un certain nombre d'exemples concrets avec des échantillons aléatoires pour illustrer l'inférence conditionnelle fondée sur un plan et les difficultés qu'elle comporte.

Robinson (1987) s'intéresse à l'inférence conditionnelle fondée sur un échantillonnage aléatoire simple lorsque seul le total de population x d'une variable concomitante x est connu. En prenant comme argument de condition la moyenne d'échantillon observée \bar{x} , il montre que l'estimateur par quotient habituel $\bar{Y}_r = (\bar{y}/\bar{x})X$ est conditionnellement biaisé. Il réussit à définir un estimateur par quotient redressé en fonction du biais conditionnel

$$(2.1) \quad Y_r(red) = \bar{Y}_r + N(r - b)(\bar{x} - \bar{X})\bar{X}/\bar{x},$$

où $r = \bar{y}/\bar{x}$ et b est le coefficient de régression de l'échantillon. Il montre aussi qu'un estimateur courant de la variance

$$(2.2) \quad s_c^2(Y_r) = N^2(1 - n/N) \sum_{i \in s} (y_i - rx_i)^2/n(n-1)$$

est conditionnellement biaisé tandis qu'un autre estimateur classique de la variance

$$(2.3) \quad s_c^2(Y_r) = (\bar{X}/\bar{x})^2 s_c^2(Y_r)$$

est en réalité conditionnellement sans biais lorsque n est grand. Robinson montre également par une étude de simulation que $s_c^2(Y_r)$ ressemble beaucoup à l'estimateur de la variance conditionnelle de $Y_r(red)$.

2.2 Approche dépendante d'un modèle

Une approche dépendante d'un modèle implique à strictement parler un échantillonnage par choix raisonné et la distribution de modèle (constituée de réalisations hypothétiques de $y = (y_1, \dots, y_N)'$ qui satisfont le modèle) permet de faire des inférences valables sur l'échantillon particulier s qui a été prélevé.

L'échantillon est un sous-ensemble s de U , avec les valeurs y correspondantes, c'est-à-dire, $\{(i, y_i), i \in s\}$, prélevé suivant un plan d'échantillonnage qui lui attribue une probabilité connue $p(s) \geq 0$ pour tous $s \in S$ (l'ensemble de tous les s possibles) et $\sum_{s \in S} p(s) = 1$. La probabilité d'échantillonnage $p(s)$ peut dépendre de variables du plan connues $z = (z_1, \dots, z_N)$, comme les variables indicatrices de strate et les tailles de grappes, c.-à-d. $p(s | z) = p(s)$ où z_j est probablement sous forme de vecteur. Dans le cas de l'échantillonnage probabiliste, les probabilités de sélection $\pi_j = \sum_{\{s: j \in s\}} p(s)$ sont positives, ce qui permet d'obtenir un estimateur non biaisé ou convergent de X au sens classique. Il est aussi d'usage de poser comme condition que les probabilités de sélection composées $\pi_{ij} = \sum_{\{s: \{i, j\} \in s\}} p(s)$ soient positives, ce qui permet d'obtenir un estimateur non biaisé ou convergent de la variance au sens classique.

La question fondamentale est de faire des inférences (estimation, estimation de variance et construction d'intervalles de confiance) à propos du total X en observant un échantillon prélevé selon un plan $p(s)$ précis tout en profitant de l'information supplémentaire disponible. L'opération comporte essentiellement trois étapes: i) choix d'un plan d'échantillonnage; ii) choix d'un estimateur X , iii) choix d'un estimateur de la variance et d'intervalles de confiance. Ces étapes peuvent être réalisées selon trois approches différentes: i) approche fondée sur un plan, que l'on appelle aussi échantillonnage probabiliste ou randomisation; ii) approche dépendante d'un modèle, aussi appelée approche prédictive ou méthode des probabilités hypothétiques (Hajek 1981); et, iii) une approche hybride, que l'on appelle approche fondée sur un modèle. Nous faisons le point ci-dessous sur chacune de ces méthodes.

2.1 Approche fondée sur un plan

Cette approche utilise la méthode probabiliste tant pour l'échantillonnage que pour l'inférence fondée sur les données. La distribution d'échantillonnage probabiliste permet, même dans des cas complexes, de faire des inférences valables quelles que soient les valeurs y de la population, en ce sens que la variable-pivot $t = (X - Y)/S(X)$ est distribuée approximativement selon une loi $N(0, 1)$, (à tout le moins pour de grands échantillons), où $S(X)$ est l'erreur type de X . On a critiqué cette approche en faisant valoir que, malgré l'absence d'hypothèses, les inférences en question portaient sur des échantillons répétés de la population sondée (ce qui comprend tous les échantillons $s \in S$ et les probabilités correspondantes $p(s)$) plutôt que sur l'échantillon particulier s qui avait été prélevé. Nous pouvons réfuter en partie ces objections en ayant recours soit à l'inférence conditionnelle fondée sur un plan, où il est question d'un sous-ensemble de S qui "a un rapport" avec l'échantillon particulier s , ou à une approche fondée sur un modèle.

Horvitz et Thompson (1952) ont contribué de façon notable à l'élaboration de la théorie de l'inférence fondée sur un plan en définissant trois catégories d'estimateurs linéaires de X puis en évoquant la possibilité que parmi tous les estimateurs linéaires non biaisés possibles de X , il ne s'en trouve aucun qui soit le meilleur estimateur (estimateur à variance minimum), même dans le cas de l'échantillonnage aléatoire simple. S'inspirant de l'article d'Horvitz et Thompson, Godambe (1955) a proposé une catégorie générale d'estimateurs linéaires définis par l'équation $X_b = \sum_{i \in s} b_{si} y_i$, où le poids b_{si} se rapporte à l'élément i si s a été prélevé et $i \in s$. Godambe a démontré que cette catégorie d'estimateurs ne comprenait pas de meilleur estimateur non biaisé de X pour aucun plan d'échantillonnage $p(s)$. Le critère de la variance minimum ayant été écarté, on en proposa plusieurs autres pour le choix d'un estimateur. Parmi ceux proposés, le critère d'admissibilité est relativement utile sauf qu'il ne permet pas de faire une distinction nette entre les avantages des divers estimateurs puisqu'un trop grand nombre d'estimateurs sont admissibles à la fois. Ghosh (1987) fait une excellente analyse des résultats observés pour le critère d'admissibilité et les critères connexes en ce qui a trait à l'échantillonnage dans une population finie. Par la même occasion, il propose de nouveaux critères pour le choix d'un seul estimateur possible parmi ceux de la catégorie de Godambe pour n'importe

que l'on puisse tirer de chaque sous-échantillon une estimation acceptable du paramètre d'intérêt. En attribuant les sous-échantillons à des interviewers différents (ou à des équipes d'interviewers différentes), on peut établir une estimation acceptable de la variance totale, qui tiennent compte de la variance de réponse corrélée due aux interviewers. Deming (1960) s'est beaucoup servi de cette méthode (appelée parfois échantillonnage répété) pour établir des estimations simples de la variance. Cela a donné naissance à des techniques de ré-échantillonnage comme la méthode jackknife, la méthode BRR (balanced repeated replication) et la méthode boots-trap, qui permettent d'estimer la variance de statistiques non linéaires complexes (voir section 3). Enfin, parmi les autres idées qui ont été formulées à propos des enquêtes à plan de sondage complexe, ne manquons pas de souligner la notion d'effet du plan (DEFF), que l'on doit à Leslie Kish (voir Kish 1965, section 8.2). L'effet du plan est défini comme le rapport entre la variance d'une statistique selon un plan d'échantillonnage donné et la variance de la même statistique selon un échantillonnage aléatoire simple (échantillons de même taille). La notion d'effet du plan s'est avérée particulièrement utile pour la présentation et la modélisation des erreurs d'échantillonnage de même que pour l'analyse de données d'enquête tirées d'échantillons en grappes et d'échantillons stratifiés (voir section 4).

2. FONDEMENTS THÉORIQUES

Bien que Neyman (1934) et d'autres aient réussi à définir des meilleurs estimateurs linéaires sans biais pour des plans de sondage simples à l'aide du modèle ordinaire de Gauss-Markov, la théorie classique des sondages s'est formée plus ou moins par induction. On considèrerait les estimateurs (et les plans) qui paraissent acceptables et on en analyserait soigneusement les caractéristiques à l'aide de méthodes analytiques ou empiriques; cette analyse prenait le plus souvent la forme d'une comparaison des biais et des erreurs quadratiques moyennes. Dans certains cas, on avait recours aussi à l'erreur quadratique moyenne ou à la variance espérée suivant des modèles de superpopulation plausibles. Comme le font remarquer Hansen et coll. (1983), on n'insistait pas sur la propriété d'être sans biais des estimateurs selon un plan de sondage donné car cette propriété "se traduit souvent par des erreurs quadratiques moyennes excessives" (traduction). On mettait plutôt l'accent sur la convergence asymptotique des estimateurs selon le plan, à tout le moins lorsqu'il fallait établir des estimations d'aggrégats à partir d'échantillons suffisamment grands, et on comparait les erreurs quadratiques moyennes de certains estimateurs asymptotiquement convergents selon le plan afin de choisir un estimateur (et un plan) acceptable. De plus, en ce qui a trait aux grandes enquêtes qui comportent de très nombreuses statistiques, on est souvent plus préoccupé d'appliquer des méthodes d'estimation uniformes que de limiter la variance de certains paramètres statistiques (comparativement à d'autres estimateurs qui sont adaptés à chaque paramètre) à cause des limites de temps et d'argent et d'autres contraintes opérationnelles.

Malgré l'utilité de l'approche classique, le besoin d'un modèle d'inférence pour données d'enquête se faisait sentir depuis longtemps. Conscients de ce besoin, plusieurs statisticiens ont contribué largement à l'élaboration de la théorie de l'inférence pour données d'enquête, surtout depuis les 10 ou 20 dernières années. Divers aspects de cette théorie sont traités dans de nombreux recueils d'articles (voir par exemple Chaudhuri, 1988) et dans deux ouvrages en particulier (Cassel et coll. 1977; Chaudhuri et Vos 1988).

La plupart des articles qui ont été écrits sur les fondements de la théorie des sondages pré-sentent un modèle quelque peu idéaliste. Une population sondée U est composée de N éléments distincts désignés par la lettre $j = 1, \dots, N$. Il est possible de connaître **précisément** la caractéristique d'intérêt y_j (probablement exprimée sous forme de vecteur) rattachée à l'élément j en observant cet élément. On suppose donc qu'il n'existe pas d'erreur de réponse ou de mesure ou bien, si elles existent, on n'en tient pas compte. Le paramètre d'intérêt est le total de population $Y = y_1 + \dots + y_N$ ou la moyenne de population $\bar{Y} = Y/N$ (si N est connu).

est venu modifier cette conception étroite de l'échantillonnage en introduisant la notion d'échantillonnage stratifié avec répartition "optimale" et d'échantillonnage en grappes avec estimation par quotient. Dans les deux cas, on peut obtenir des estimations "valables" de totaux, de moyennes ou de proportions de population sans devoir s'appuyer sur un échantillon représentatif prélevé à l'aide d'un plan avec probabilités de sélection égales. Enfin, Neyman a aussi fait avancer la théorie des sondages en introduisant la notion de fonction de coût (Neyman 1938); cette fonction permet de déterminer, dans un sondage à deux phases, le mode de répartition de l'échantillon pour lequel la variance est minimum, étant donné un budget défini.

Les travaux de Neyman ont amené d'autres statisticiens à poursuivre la recherche dans le même sens et à compléter, le cas échéant, l'oeuvre de Neyman. Ainsi, parmi les principaux sujets de recherche, il convient de mentionner l'estimation par quotient et par régression dans le sondage à deux phases (Cochran 1939), la détermination de points de stratification "optimale" et du mode de répartition "optimale" en présence de plusieurs paramètres ou caractéristiques (Dalenius 1957), et l'échantillonnage répété une fois avec remise partielle des unités (Jessen 1942); ce dernier sujet a été approfondi par Patterson (1950) et Hansen et coll. (1953, p. 470-503), qui en sont venus à parler d'échantillonnage répété plus d'une fois (ou d'échantillonnage avec renouvellement). L'échantillonnage avec renouvellement et les estimateurs "composites" qui s'y rattachent servent souvent aujourd'hui à estimer des niveaux et des variations à l'aide de données tirées des grandes enquêtes permanentes à objectifs multiples (par exemple la Current Population Survey (CPS), réalisée par le U.S. Bureau of the Census).

Les travaux de Neyman ont eu aussi une influence déterminante sur les recherches qu'effectuaient Morris Hansen, William Hurwitz et leurs collègues au U.S. Bureau of the Census. Situés par les difficultés que posaient les plans de sondage des grandes enquêtes et par la façon dont Neyman abordait la théorie des sondages, Hansen et Hurwitz (1943) ont élaboré la théorie de l'échantillonnage avec probabilité proportionnelle à la taille avec remise (aussi appelé échantillonnage PPT). Dans les enquêtes à plusieurs degrés, cette méthode permet de répartir à peu près également la charge de travail entre les interviewers, ce qui facilite la gestion de ce genre d'enquêtes. De plus, l'échantillonnage PPT contribue à réduire sensiblement la variance des estimations en limitant la variabilité attribuable à l'inégalité des tailles de grappe sans que soit nécessaire une stratification selon la taille et en favorisant ainsi la stratification selon d'autres variables de manière à réduire la variance. Hurwitz et Thompson (1952) et Narain (1951) ont poussé plus loin les recherches de Hansen et Hurwitz et en sont venus à l'échantillonnage avec probabilités inégales sans remise. En posant la probabilité de sélection de chaque unité proportionnelle à sa taille à chaque degré d'échantillonnage, on conserve les aspects intéressants de la méthode de Hansen-Hurwitz; on utilise en l'occurrence l'estimateur d'Hurwitz-Thompson d'un total de population. Les travaux de Horvitz et Thompson et de Narain ont ouvert la voie à de nombreuses études théoriques et pratiques sur l'échantillonnage avec probabilités inégales sans remise. Brewer et Hanif (1983) et Chaudhuri et Vos (1988) font un compte-rendu détaillé de ces études.

Madaw et Madaw (1944) ont jeté les bases de la théorie de l'échantillonnage systématique et ont défini des modèles de population pour analyser les caractéristiques de ce type d'échantillonnage. Cochran (1946) a élaboré la notion de "superpopulation", selon laquelle la population finie est tirée d'une superpopulation infinie qui présente certaines caractéristiques. Un modèle de superpopulation permet d'établir la valeur espérée (ou prévue) de variances; on compare ensuite ces variances entre elles afin d'analyser l'efficacité relative de diverses méthodes d'échantillonnage. L'article de Cochran (1946) est à l'origine de nombreuses autres études sur l'utilisation des modèles de superpopulation dans le choix des méthodes d'échantillonnage et sur l'importance dépendante d'un modèle ou l'inférence fondée sur un modèle (voir section 2). Mahalanobis (1946) a élaboré une méthode d'interprétation des sous-échantillons et l'a utilisée largement dans de grandes enquêtes en Inde afin d'évaluer les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage. La méthode consiste à prélever au moins deux sous-échantillons indépendants (d'un même échantillon) selon le même plan de sondage de sorte

Genèse et évolution des fondements théoriques de l'estimation et de l'analyse fondées sur les sondages

J.N.K. RAO et D.R. BELLHOUSE¹

RÉSUMÉ

À l'origine, les recherches en théorie et pratique des sondages étaient surtout orientées vers l'élaboration de plans de sondage efficaces et de méthodes d'estimation de total ou de moyenne de population tout aussi efficaces. Par la suite, on s'est attaché à faire une analyse critique des fondements théoriques de l'estimation fondée sur les sondages et on a proposé des modèles d'inférence pour les totaux ou les moyennes. Durant les dix dernières années, des progrès sensibles ont été réalisés dans l'élaboration de méthodes d'analyse de données d'enquête qui tiennent compte de la complexité du plan de sondage. Dans cet article, nous passons en revue quelques-unes des étapes de cette évolution et nous en faisons l'évaluation.

MOTS CLÉS: Fondements de l'inférence; analyse de données d'enquête; logiciel.

1. LES PREMIERS JALONS DE LA RECHERCHE EN SONDAGES

Avant 1950 ou même 1960, la recherche sur les enquêtes par sondage était surtout motivée par le désir d'obtenir, au coût voulu, des estimateurs de totaux, de moyennes ou de proportions raisonnablement efficaces pour de grandes populations de plus en plus complexes. De nombreux recueils d'articles renferment des premières études qui ont été faites sur l'échantillonnage de populations humaines (voir par exemple Hansen, Dalenius et Tepping

1985 et Bellhouse 1988).

La théorie mathématique des sondages remonte à la fin du dix-neuvième siècle, à l'époque du statisticien norvégien A.N. Kiaer. Celui-ci fut le premier à faire valoir les mérites de ce qu'on appelait à l'époque "la méthode représentative", ou l'échantillonnage, par rapport au dénombrement complet. Pour Kiaer (1897), un échantillon représentatif était un échantillon qui devait refléter la population finie d'où il était tiré. Il y avait deux façons d'obtenir un tel échantillon: par randomisation ou par échantillonnage raisonné (équilibre). À l'origine, la seconde méthode avait la faveur de tous mais peu à peu, la randomisation devint une sérieuse alternative. Dans les années 1920, l'une et l'autre étaient largement utilisées. Bowley (1926) fait un résumé des principales étapes qui ont marqué, à cette époque, l'évolution des deux méthodes d'échantillonnage. Il signale notamment la création de l'échantillonnage aléatoire stratifié avec répartition proportionnelle et l'élaboration de formules permettant de déterminer le degré de précision d'une estimation tirée d'un échantillon choisi de façon raisonnée.

La publication du fameux article de Neyman (1934) est venue défaire le rapport d'égalité qui existait entre l'échantillonnage aléatoire et l'échantillonnage par choix raisonné. Neyman avait réussi à expliquer, tant du point de vue théorique que pratique, pourquoi la première méthode devait être préférée à la seconde dans les grandes enquêtes qui avaient cours à l'époque. De plus, l'article de Neyman a favorisé un approfondissement des méthodes d'échantillonnage aléatoire. Auparavant, Bowley et ses collaborateurs n'utilisaient que des plans de sondage qui prévoyaient la même probabilité de sélection pour toutes les unités de la population. C'était pour eux la seule façon d'obtenir un échantillon représentatif. Neyman (1934)

¹ J.N.K. Rao, Département de mathématique et de statistique, Université Carleton, Ottawa (Ontario), K1S 5B6.
D.R. Bellhouse, Département de statistique, Université Western Ontario, London (Ontario), N6A 5B9.

la résistance à laquelle s'est heurtée l'introduction du principe de l'échantillonnage, qui aujourd'hui nous paraît aller de soi. Ses commentaires pénétrants sur diverses questions particulières sont trop nombreux pour que nous puissions les résumer ici.

Le propos de Dalenius et de Särndal était à l'origine de faire un commentaire sur l'article de Bailar, mais leur article s'est transformé en un exposé sur l'histoire des méthodes des sondages en Suède. Dans la forme qu'il a prise, cet exposé peut servir de résumé et de mise à jour du livre publié par Dalenius en 1957.

Les autres articles de ce numéro de Techniques d'enquête traitent de sujets divers. Kott propose un estimateur de variance sans biais pour un plan de sondage à deux phases où il y a échantillonnage aléatoire simple stratifié dans chacune des deux phases. Ce plan de sondage est d'usage courant, en particulier dans les enquêtes sur l'agriculture.

L'échantillonnage à deux phases avec stratification aux deux phases est également le sujet de l'article de White. L'auteur étudie au moyen de la simulation un estimateur mis au point par Vardeman et Meeden et qui utilise l'information préalable. Il donne également des résultats théoriques pour le cas où l'information préalable n'est pas utilisée.

Julien et Maranda décrivent le plan de sondage utilisé depuis 1988 pour l'enquête nationale sur les fermes. Les auteurs évaluent l'efficacité du nouveau plan en comparant le degré de précision des estimations de 1988 à celui des estimations de 1987 de même qu'au degré de précision attendu qui avait été déterminé au stade de la mise au point du nouveau plan.

Dans son article, Hay cherche à déterminer les effets de la méthode de collecte de données sur les réponses en analysant les résultats d'une étude effectuée en Saskatchewan. Il compare le questionnaire à remplir soi-même à l'interview sur place. Il y a des différences statistiquement significatives, mais pas suffisantes pour avoir une importance pratique.

Langlet étudie l'utilisation de l'analyse typologique comme moyen de résoudre le problème de l'imputation en cas de non-réponse partielle. Cette technique serait particulièrement utile dans les cas où le nombre de classes d'imputation est plutôt élevé.

Béland et Théberge font appel aux tests de randomisation pour comparer deux questionnaires utilisés pour étudier les questions susceptibles d'être posées lors du recensement de 1991. Comme les tests de ce genre peuvent n'être pas bien connus des méthodologistes d'enquête, cet article pourra constituer une introduction utile.

Dans son article, Cantwell établit une formule de variance simple pour un estimateur composite général souvent employé dans les plans de sondage avec renouvellement. L'auteur considère ces plans avec un seul ou avec plusieurs niveaux.

Le rédacteur en chef

Dans ce numéro

Dans la section spéciale de ce numéro, nous considérons le passé et nous regardons vers l'avenir. Les auteurs des articles figurant dans cette section sont des statisticiens connus, spécialisés dans la conception d'enquêtes; le lecteur pourra tirer profit de la richesse de leurs connaissances et de leur expérience. En jetant un regard éclairant sur les progrès réalisés en matière de techniques d'enquête, ces collaborateurs nous donnent la possibilité de nous intéresser aux domaines où se font les recherches les plus prometteuses. À l'exception d'un seul, chaque article est suivi de commentaires et d'une réponse de l'auteur à ces commentaires.

Rao et Bellhousse présentent un aperçu historique de la théorie et des méthodes des sondages. Après un exposé sur les premiers acquis dans ce domaine, les auteurs nous parlent de la conception plan/modèle, des méthodes d'estimation de la variance, de l'analyse des données d'enquête et des développements récents en matière de logiciels. L'article contient une importante bibliographie. Les commentaires de Smith complètent l'article par une perspective un peu différente sur ces questions et par des réflexions sur le statut de la théorie des sondages par rapport à la statistique traditionnelle.

Après un exposé du rôle que l'Etat et les chercheurs ont eu dans les débuts des enquêtes par sondage et des recensements, Fienberg et Tanur décrivent les bases institutionnelles des sondages, particulièrement aux Etats-Unis. Les organismes d'Etat, les associations de statisticiens, les maisons de sondage et les universités sont au nombre des organisations considérées. Les auteurs étudient les progrès récents, dont le recours accru à l'interview téléphonique et les aspects cognitifs des sondages. Ils terminent leur article en décrivant les liens qui existent entre les différents groupes d'institutions actives dans le domaine des enquêtes. Dans ses commentaires, Groves considère également ces groupes d'institutions et fait remarquer que le passage de chercheurs de l'un à l'autre a été moins fréquent que ne le laissent entendre Fienberg et Tanur. Groves allonge aussi considérablement la liste des progrès récents.

Tandis que Fienberg et Tanur considèrent les institutions gouvernementales comme un élément parmi d'autres, Bailar s'attache à l'importance du rôle du U.S. Bureau of the Census dans le développement des méthodes des sondages. Elle montre pourquoi et comment on a mis au point les différentes méthodes et approches, y compris l'échantillonnage et la désaisonnalisation. L'article s'achève sur une perspective ouverte sur l'avenir. Brackstone souligne que ce sont des problèmes pratiques qui ont engendré les progrès dont parle Bailar. Il mentionne plusieurs autres à celles dont parle Bailar. Brackstone souligne aussi l'importance d'un climat propice pour favoriser l'innovation. Dans son article, Kish parle de formules qui pourraient remplacer l'actuel recensement périodique. Il ranime le débat sur la possibilité de le remplacer par le recensement par étapes. Il considère l'utilisation des données administratives à cette fin, en soulignant qu'il existe de bonnes sources de données de ce genre dans certains pays. La manière d'accumuler les données obtenues au moyen de sondages et de recensements par étapes est un point important. Diverses solutions sont envisagées. Dans ses commentaires, Scheuren fait remarquer que Kish, en réalité, propose de changer en profondeur notre manière de penser, tâche toujours difficile. Bien qu'il estime que le recensement par étapes comme tel serait sans doute trop coûteux, Scheuren pense qu'une formule modifiée de ce type de recensement, assortie de meilleures données administratives, serait acceptable. Kish comme Scheuren conviennent qu'il faudra encore beaucoup de recherche pour qu'il soit permis d'espérer des progrès dans ce domaine. Nous sommes très contents d'avoir pour tous les articles précédemment mentionnés les commentaires de Morris Hansen, qui a joué un rôle actif dans beaucoup des développements évoqués par les auteurs de ces articles. Il apporte des précisions historiques précieuses et rectifie certaines erreurs et certaines idées fausses. Un point particulièrement intéressant dont parle Hansen, c'est

TABLE DES MATIÈRES - fin

C. JULIEN et F. MARANDA	127
Le plan de sondage de l'enquête nationale sur les fermes de 1988.....	
D.A. HAY	141
Le choix de la méthode est-il important pour les sujets d'enquête délicats?	
E.R. LANGLET	147
Analyse typologique appliquée au regroupement de classes d'imputation	
Y. BÉLAND et A. THÉBERGE	155
Un exemple d'utilisation de tests aléatoires pour les essais du questionnaire du recensement	
P.J. CANTWELL	163
Formules de variance pour estimateurs composites dans les plans de renouvellement	

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 16, numéro 1, juin 1990

TABLE DES MATIÈRES

1	Dans ce numéro
	Section spéciale – Histoire et questions actuelles dans le domaine des recensements et des sondages
	J.N.K. RAO et D.R. BELLHOUSE
3	Genèse et évolution des fondements théoriques de l'estimation et de l'analyse fondées sur les sondages.....
27	Commentaires: T.M.F. SMITH.....
	S.E. FIENBERG et J.M. TANUR
33	Origine institutionnelle des enquêtes par sondage aux Etats-Unis: Une perspective historique
50	Commentaires: R.M. GROVES.....
	B.A. BAILAR
	Rôle de l'administration fédérale dans le développement des méthodes statistiques aux Etats-Unis.....
55	Commentaires: G.J. BRACKSTONE
	L. KISH
67	Recensement par étapes et échantillons avec renouvellement complet
78	Commentaires: F. SCHEUREN.....
	M.H. HANSEN
	Commentaires sur les articles de la section spéciale.....
87	RÉPONSES: J.N.K. RAO et D.R. BELLHOUSE
93	S.E. FIENBERG et J.M. TANUR.....
95	B.A. BAILAR.....
97	L. KISH
99	T. DALENIUS et C.-E. SÄRNDAAL
	Certains progrès relatifs aux techniques de sondage et à leur utilisation dans les statistiques officielles en Suède
101	
	P.S. KOTT
	Estimation de la variance lorsque l'échantillon aréolaire de première phase est restreint
107	
	D.B. WHITE
113	Estimation au moyen d'un échantillonnage double et d'une stratification duale

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

B. Afonja, *Nations Unies*

D.R. Bellhouse, *U. of Western Ontario*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

J.C. Deville, *INSEE*

D. Drew, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of*

Management and Budget

Rédacteurs adjoints

J. Gambino, L. Mach et A. Thèberge, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M. P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, 4^e étage, Edifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 30 \$ par année au Canada, 36 \$ (É.-U.) aux États-Unis, et de 42 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA
JUIN 1990

Publication autorisée par le ministre
de l'Industrie, des Sciences et de la Technologie
©Ministre des Approvisionnements
et Services Canada 1990

Tous droits réservés. Il est interdit de reproduire ou
de transmettre le contenu de la présente publication,
sous quelque forme ou par quelque moyen que ce soit,
enregistré ou sur support magnétique, reproduction
électronique, mécanique, photographique, ou autre, ou
de l'emmagasiner dans un système de recouvrement,
sans l'autorisation écrite préalable du ministre
des Approvisionnements et Services Canada

Septembre 1990

Prix: Canada: 30 \$ par année
États-Unis: 36 \$ US par année
Autres pays: 42 \$ US par année

Catalogue 12-001, vol. 16, n° 1
ISSN 0714-0045

Ottawa

Canada

VOLUME 16, NUMÉRO 1
JUN 1990

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE

Statistics
Canada

Statistique
Canada



Catalogue 12-001



Survey Methodology

A Journal of Statistics Canada

December 1990 Volume 16 Number 2



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

December 1990 Volume 16 Number 2

Published under the authority of the Minister
of Industry, Science and Technology

© Minister of Supply and Services Canada 1991

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the Minister of Supply and Services Canada.

March 1991

Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

B. Afonja, <i>United Nations</i>	R.M. Groves, <i>U.S. Bureau of the Census</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	D. Holt, <i>University of Southampton</i>
D. Binder, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
E.B. Dagum, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
J.-C. Deville, <i>INSEE</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
W.A. Fuller, <i>Iowa State University</i>	C.E. Särndal, <i>University of Montreal</i>
J.F. Gentleman, <i>Statistics Canada</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

J. Gambino, L. Mach and A. Thériège, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

**STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION**

The proceedings of Symposium 88: Analysis of Data in Time are now available. The Proceedings from Symposium 90: Measurement and Improvement of Data Quality will be available in the summer of 1991. A limited number of copies of back issues from the 1987 and 1988 symposia are also available. To order, send this form to:

SYMPOSIUM PROCEEDINGS
STATISTICS CANADA
R.H. COATS BUILDING, 11TH FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada - Symposium Proceedings").

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	_____	@ \$10 EACH
1987 -	Statistical Uses of Administrative Data - FRENCH	_____	@ \$10 EACH
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	_____	@ \$12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	_____	@ \$10 EACH
1989 -	Analysis of Data in Time - BILINGUAL	_____	@ \$20 EACH
Forthcoming (advance price):			
1990 -	Measurement and Improvement of Data Quality - ENGLISH	_____	@ \$20 EACH
1990 -	Measurement and Improvement of Data Quality - FRENCH	_____	@ \$20 EACH
1990 -	SET OF 1 ENGLISH AND 1 FRENCH	_____	@ \$35 PER SET

PLEASE ADD \$2 PER VOLUME FOR SHIPPING \$ _____
TOTAL AMOUNT OF ORDER \$ _____

HIGHLIGHTS OF THE 1989 PROCEEDINGS: Analysis of Repeated Surveys (W.A. Fuller), Unique Features and Problems of Rolling Samples (L. Kish), A Time Series Model for Estimating Housing Price Indexes Adjusted for Changes in Quality (D. Pfeffermann, L. Burke, S. Ben-Tuvia), Analysis of Seasonal ARIMA Models from Survey Data (D.A. Binder, J.P. Dick), Small Area Estimation Using Models that Combine Time Series and Cross-Sectional Data (G.H. Choudhry, J.N.K. Rao), Mapping Aggregate Birth Data (D.R. Brillinger), Analysis of Cross-Classified Categorical Time Series (A.C. Singh, G.R. Roberts), Alternative Approaches to the Analysis of Time Series Components (W.R. Bell, M.G. Pugh), Adjustment for Reporting-Delay of AIDS and Estimation of the Size of the HIV Infected Population in the U.S.A. (I.B. MacNeill, Q.P. Duong, V.R. Jandhyala, L. Lu), Some Statistical Methods for Panel Life History Data (J.D. Kalbfleisch, J.F. Lawless).

HIGHLIGHTS OF THE 1990 PROCEEDINGS: Managing Quality in National Statistics Programs (J. Early), Sampling Flows of Mobile Human Populations (G. Kalton), Techniques to Control and Improve the Quality of Data in Large Databases (T.C. Redman, R.W. Pauke), Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used (C.E. Sarndal), A Systematic Approach to Quantifying the Quality of Repeated Surveys (R. Tortora), A Review of Some Macroediting Methods for Rationalizing the Editing Process (L. Granquist), The Measurement of Net Coverage Error in Canadian Censuses (R.G. Carter), Differential Coverage in the United States Census of Population (G. Robinson, H. Hogan), Measurement of Content Data Quality in the 1990 Census (H. Wolman, K.F. Thomas).

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME _____
ADDRESS _____
CITY _____ PROV/STATE _____ COUNTRY _____
POSTAL CODE _____ TELEPHONE (____) _____ FAX _____

Please note: Each Symposium registrant not employed by Statistics Canada receives one free copy of the Proceedings

LA SÉRIE DES SYMPOSIUMS INTERNATIONAUX DE STATISTIQUE CANADA
RENSEIGNEMENTS CONCERNANT LA COMMANDE DES RECUEILS

Le recueil du Symposium 89: "L'analyse des données dans le temps" est maintenant disponible. Le recueil du Symposium 90: "Mesure et amélioration de la qualité des données" sera disponible à l'été 1991. Un nombre limité des copies des recueils des symposiums 1987 et 1988 sont aussi disponibles. Pour commander, envoyez cette formule à l'adresse suivante:

RECUEIL DU SYMPOSIUM
STATISTIQUE CANADA
ÉDIFICE R.H. COATS, 11^e ÉTAGE
PARC TUNNEY
OTTAWA (ONTARIO) CANADA
K1A 0T6

Veuillez inclure le paiement avec votre commande (chèque ou mandat, en dollars canadiens ou l'équivalent, payable à "Receveur général du Canada - Recueil du Symposium").

RECUEIL DU SYMPOSIUM: NUMEROS DISPONIBLES

1987 - Les utilisations statistiques des données administratives - ANGLAIS	_____ @ \$10 CHACUN
1987 - Les utilisations statistiques des données administratives - FRANÇAIS	_____ @ \$10 CHACUN
1987 - ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	_____ @ \$12 L'ENSEMBLE
1988 - Les répercussions de la technologie de pointe sur les enquêtes - BILINGUE	_____ @ \$10 CHACUN
1989 - L'analyse des données dans le temps - BILINGUE	_____ @ \$20 CHACUN

Disponible bientôt (prix de lancement):

1990 - Mesure et amélioration de la qualité des données - ANGLAIS	_____ @ \$20 CHACUN
1990 - Mesure et amélioration de la qualité des données - FRANÇAIS	_____ @ \$20 CHACUN
1990 - ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	_____ @ \$35 L'ENSEMBLE

S.V.P. AJOUTEZ \$2 PAR LIVRE POUR LES FRAIS DE LIVRAISON \$ _____

MONTANT TOTAL DE LA COMMANDE \$ _____

POINTS CULMINANTS DU RECUEIL 1989: Analyse d'enquêtes à passages répétées (W.A. Fuller), Caractéristiques et problèmes propres aux échantillons successifs (L. Kish), Modèle de série chronologique ajusté pour tenir compte des variations de qualité et servant à l'estimation des indices de prix du logement (D. Pfeffermann, L. Burke, S. Ben-Tuvia), Analyse des modèles ARMMI saisonniers au moyen des données d'enquête (D.A. Binder, J.P. Dick), Estimation des données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales (G.H. Choudhry, J.N.K. Rao), Représentation cartographique de données agrégées (D.R. Brillinger), Analyse de séries chronologiques qualitatives en tableaux croisés (A.C. Singh, G.R. Roberts), Autres approches d'analyse des éléments des séries chronologiques (W.R. Bell, M.G. Pugh), Ajustement pour tenir compte des déclarations tardives des cas de SIDA et estimation de la population infectée par le VIH aux États-Unis (I.B. MacNeill, Q.P. Duong, V.R. Jandhyala, L. Liu), Quelques méthodes statistiques d'analyse de données historiques personnelles de panel (J.D. Kalbfleisch, J.F. Lawless).

POINTS CULMINANTS DU RECUEIL 1990: La gestion de la qualité dans les programmes statistiques nationaux (J. Early), L'échantillonnage des flux de populations humaines mobiles (G. Kalton), Techniques pour contrôler et améliorer la qualité des données des grandes bases de données (T.C. Redman, R.W. Pautke), Méthodes pour estimer la précision des estimations d'enquêtes lorsqu'il y a eu imputation (C.E. Särndal), A Systematic Approach to Quantifying the Quality of Repeated Surveys (R. Tortora), Une revue de certaines méthodes de macro-vérification pour rationaliser le processus de vérification (L. Granquist), Évaluation de l'erreur de couverture nette dans les recensements Canadiens (R.G. Carter), Couverture différentielle du recensement de la population aux États-Unis (G. Robinson, H. Hogan), Mesure de la qualité des données du recensement de 1990 (H. Woltman, K.F. Thomas).

S.V.P. INCLURE VOTRE ADRESSE POSTALE COMPLÈTE AVEC VOTRE COMMANDE

NOM _____

ADRESSE _____

VILLE _____ PROV/ÉTAT _____ PAYS _____

CODE POSTAL _____ TÉLÉPHONE (____) _____ FAX _____

Prétez attention s.v.p.: Chaque participant au Symposium qui n'est pas un employé de Statistique Canada recevra une copie gratuite du recueil du Symposium.

MORRIS H. HANSEN
(1910-1990)

This issue is dedicated to the memory of Morris H. Hansen,
a pioneer, innovator and leader
who made fundamental and lasting contributions
to many aspects of survey methodology.

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 16, Number 2, December 1990

CONTENTS

In This Issue	165
Time Series Methods in Surveys	
W.A. FULLER	
Analysis of Repeated Surveys	167
K.M. WOLTER and R.M. HARTER	
Sample Maintenance Based on Peano Keys	181
W.R. BELL and S.C. HILLMER	
The Time Series Approach to Estimation for Repeated Surveys	195
D. PFEFFERMANN and L. BURCK	
Robust Small Area Estimation Combining Time Series and Cross-Sectional Data ..	217
D.A. BINDER and J.P. DICK	
A Method for the Analysis of Seasonal ARIMA Models	239
D.R. BRILLINGER	
Spatial-Temporal Modelling of Spatially Aggregate Birth Data	255
N. LANIEL and K. FYFE	
Benchmarking of Economic Time Series	271
<hr/>	
S. BANDYOPADHYAY	
Forgot the Sampling Scheme at the Estimation Stage?	279
H. LEE	
Estimation of Panel Correlations for the Canadian Labour Force Survey	283
A.R. SILBERSTEIN	
First Wave Effects in the U.S. Consumer Expenditure Interview Survey	293
E.A. STASNY	
Symmetry in Flows Among Reported Victimization Classifications with Nonresponse	305
Acknowledgements	331

In This Issue

This issue contains a special section on time series methods in surveys, a topic that has attracted considerable interest in recent years. Special thanks are due to W.A. Fuller and J.N.K. Rao for coordinating the editorial work for this section.

The first two papers of the special section deal with the problems of sample design and maintenance, and estimation of various parameters of interest in repeated surveys. Fuller notes that repeated surveys designed to enable estimation of the parameters of the measurement error process can be very cost efficient. For a two-period survey with fifty percent overlap, he shows that generalized least square estimates of longitudinal parameters can have substantially lower variance than the simple estimator based only on the overlapping units. Wolter and Harter deal with the problem of sample maintenance for a recurring survey. The ingenious use of a Peano curve allows the sample maintenance to meet several desirable properties. They describe an application to a marketing survey.

Bell and Hillmer discuss the underlying philosophy of the time series approach to estimation in repeated surveys based on the recognition of two sources of variation: time series variation and sampling variation. They obtain some theoretical results regarding design consistency of the time series estimators, and uncorrelatedness of the signal and sampling error series. They also observe that the use of signal extraction results from time series analysis can improve survey estimates by reducing their mean square error.

For repeated surveys, better small area estimates can be obtained by combining the usual approach based on synthetic estimation with the use of time series models. Pfeiffermann and Burck examine the statistical properties of such predictors. They illustrate the procedure with the use of data on home sale prices.

Time series described by ARIMA regression models with survey errors following an ARMA process is the subject of Binder and Dick's paper. Such models can be applied to data from surveys with a two-stage design where the first stage units are replaced randomly, while the second stage units have a rotating panel design. The authors give an example using Labour Force Survey data.

Brillinger studies the relationship of births to time and geography using data for women aged 25-29 in Saskatchewan. Smooth surfaces are obtained from data aggregated by census division. The Poisson-lognormal distribution is also fitted to the data.

In the last paper of the special section, Laniel and Fyfe describe the problem of benchmarking sub-annual series and briefly review some solutions proposed in the literature. They then present two new methods – one based on a model for trends and the other on a model for levels – and discuss their suitability.

In his paper, Bandyopadhyay proves that for a class of estimators and sampling schemes, one can ignore the sampling weights when estimating a ratio. He applies this to a well-known example to illustrate the result and makes a comparison with estimation using a ratio of Horvitz-Thompson estimators.

In repeated surveys with rotation panels, knowledge of panel correlations is essential for certain statistical analyses, such as studies of composite estimators. Lee provides methodology for estimating correlations between panel estimates in the Canadian Labour Force Survey.

Misdating or "telescoping" is a recognized source of errors in retrospective surveys. Silberstein estimates telescoping effects to obtain estimates for the unbounded first wave in the U.S. Consumer Expenditure Interview Survey. She finds that estimates from the first wave are greater than estimates from subsequent waves even after accounting for telescoping effects and concludes that a shorter recall period for the first wave improves reporting in subsequent waves.

Stasny presents several models for gross flows in the presence of nonresponse. The models are divided into those with symmetric and asymmetric transition probabilities. Methods for obtaining parameter estimates for the various models are developed and applied to victimization data from the U.S. National Crime Survey.

Finally, readers will notice that, with this issue, *Survey Methodology* has a new cover. The previous cover was used since December 1984 (Vol. 10 No. 2). Statistics Canada is making similar changes to all its publications to incorporate a unique logo and to create a standardized corporate look.

The Editor

Analysis of Repeated Surveys

WAYNE A. FULLER¹

ABSTRACT

Repeated surveys in which a portion of the units are observed at more than one time point and some units are not observed at some time points are of primary interest. Least squares estimation for such surveys is reviewed. Included in the discussion are estimation procedures in which existing estimates are not revised when new data become available. Also considered are techniques for the estimation of longitudinal parameters, such as gross change tables. Estimation for a repeated survey of land use conducted by the U.S. Soil Conservation Service is described. The effects of measurement error on gross change estimates is illustrated and it is shown that survey designs constructed to enable estimation of the parameters of the measurement error process can be very efficient.

KEY WORDS: Survey sampling; Least squares; Measurement error; Gross change.

1. INTRODUCTION

There is considerable interest in the analysis of surveys that are repeated in time. Evidence of this interest is the recently published proceedings of a conference on panel surveys edited by Kasprzyk, Duncan, Kalton and Singh (1989), sessions at the meetings of the International Statistical Institute held in 1987 and 1989, and the Statistics Canada Symposium on Analysis of Data in Time held in October 1989. Smith and Holt (1989) at the 1989 ISI session in Paris call this a "resurgence of interest in the design and analysis of longitudinal studies." They note that researchers in areas such as sociology and health have long conducted panel surveys and cohort studies. They cite, as an example, Lazarsfeld and Fiske (1938). An example in a health related area is the study of Garcia, Battese, and Brewer (1975).

Official agencies conduct many surveys, such as labor force surveys, on a regular basis. The output of such surveys is usually a sequence of reports, such as those on current employment and unemployment. Typically, very few statistics on the behavior of individual units over time have been reported from repeated official surveys. An example of a survey designed to produce longitudinal estimates is the U.S. Survey of Income and Program Participation. See Kasprzyk and McMillen (1987). While information on private surveys is less complete than that on government surveys, it seems that the most common use of repeated private surveys is also to produce a sequence of reports for points in time. However, the demand for longitudinal analysis has increased for both public and private data providers.

The complex issues associated with repeated surveys are brought into focus when one attempts to develop a taxonomy for such studies. Duncan and Kalton (1987) list some seven objectives of surveys repeated over time. These are:

- A. To provide estimates of population parameters at distinct time points.
- B. To provide estimates of population parameters summed across time.
- C. To measure net change at the aggregate level.

¹ Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.

- D. To measure components of change including
 - i) gross change
 - ii) change for an individual
 - iii) variability for an individual.
- E. To aggregate individual data over time.
- F. To measure the frequency, timing and duration of events.
- G. To accumulate information on rare populations.

While not mentioned explicitly, several of these objectives implicitly include the estimation of the parameters of subject matter models.

Duncan and Kalton also define four kinds of surveys. Their definitions were: (1) repeated survey, in which no attempt is made to guarantee that particular elements appear in more than one sample; (2) the pure panel survey, in which the same elements are observed at every point in time; (3) the rotating panel survey, in which there is a fixed pattern under which elements are observed for a fixed number of times and then rotated out of the sample; and (4) the split panel survey, in which a pure panel survey is combined with a repeated survey or a rotating panel survey. Duncan and Kalton present a table in which they outline how the different kinds of surveys are appropriate for the different kinds of objectives.

An institution conducting a repeated survey faces all of the usual survey problems, but the problems are magnified relative to a one-time survey. The quality repetition of a survey requires maintaining consistent field, processing, data management, and estimation procedures over time. It is difficult to maintain cooperation over time and it is difficult to trace people who move. Response error is present in all surveys, but repeated surveys encounter problems of "conditioning" associated with repeated interviews. Also, response errors introduce inconsistencies into data collected over time. Finally, the changing composition of units, such as families, over time complicates estimation and analysis.

We shall examine only a few issues associated with repeated surveys. Our discussion is motivated by a large scale survey conducted by the U.S. Soil Conservation Service with the cooperation of Iowa State University. In Section 2 we review some of the estimation techniques applicable for repeated surveys. This discussion is continued in Section 3 with more emphasis on estimation of longitudinal parameters in panel surveys. In Section 4 we briefly describe the estimation procedures used in the U.S. Soil Conservation Service study. Section 5 contains a short description of the effects of measurement error on gross change estimates.

2. ESTIMATION

In this section we outline generalized least square estimation for surveys with only a subset of elements observed at successive times. Generalized least squares was the procedure first considered by authors studying estimation for surveys repeated in time. Beginning with Jessen (1942), who was influenced by Cochran (1942), these authors considered the construction of minimum variance weights for a set of unbiased estimators available at each point in time of the survey.

Jessen (1942) investigated the special case of sampling on two occasions with unequal numbers of observations, and studied the optimal allocation of units to overlapping and nonoverlapping sample groups. Patterson (1950) considered sampling on T occasions under several schemes of partial replacement of units. The simplest such sampling plan required the replacement of a fixed proportion of sampling units on each successive sampling occasion.

Also, Patterson (1950) assumed that for a given i , the differences $x_{ti} - x_t, t = 1, 2, \dots$, followed a first-order autoregressive process, where x_{ti} was the value of the i -th population unit at time t , and x_t was the corresponding finite population mean. Under the resulting error model, he developed optimal estimators of the fixed x_t values and of the differences $x_t - x_{t-1}$. He also considered the optimal estimation of x_t under generalizations of the partial replacement plan, optimal sample size selection, and estimation with nonautoregressive errors.

Least squares procedures were considered further by Eckler (1955), Gurney and Daly (1965), and Jones (1980). Composite estimation was a name given to certain types of estimators. See Rao and Graham (1964), Graham (1973) and Wolter (1979). Battese, Hasabelnaby and Fuller (1989) describe the application of the least squares procedure to a farm survey conducted by the U.S. Department of Agriculture.

It seems fair to say that the parameters under consideration by these authors were means or totals at specific time points. That is, longitudinal parameters, such as the fraction of individuals in a particular class at both time 1 and time 2, were not explicitly considered by these authors. However, as we shall see, the least squares method extends to longitudinal parameters.

Linear least squares has the desirable feature that estimators for a number of characteristics are internally consistent. That is, the least squares estimator of Y plus the least squares estimator of Z is the least squares estimator of $Y + Z$. However, if different vectors of observations are used to construct different estimates, the internal consistency is destroyed.

In many applied surveys it is not possible to compute the optimum least squares estimators for all points in time because all available information cannot be used in the estimation. First, it is not possible to incorporate all data from the surveys of preceding times into a least squares analysis for the current time because the number of variables often exceeds the number of observations. Second, the releasing organization may be restricted in the number of times they can revise previous estimates. This second point has been discussed by Smith and Holt (1989).

To illustrate these estimation problems, we have constructed a small example. A two-way table for classification at two points in time, as observed in a very large sample, is given in Table 1. We have given names to the categories in this table, letting the first category be employed and letting the second category be unemployed. We shall assume that the population is constant over time. If there are births and deaths, then the table would need to be increased to a 3×3 table. Let us assume that we are interested in estimating the change in level from one period to the next. Let us also assume that we are interested in the gross change table which involves estimating the interior cells of the table. In the 2×2 table it is only necessary to estimate the (1, 1) cell and the marginal proportions to define all cells of the table.

We assume a two-period study in which an equal number of elements are observed at each of the two times. We assume that one half of the elements observed at the first time are also observed at the second time. That is, of the elements observed at the second time, one half

Table 1
Hypothetical proportions for two points in time

TIME 1	TIME 2		
	Employed	Unemployed	Total
Employed	0.91	0.02	0.93
Unemployed	0.03	0.04	0.07
Total	0.94	0.06	1.00

Table 2
Covariance matrix of the vector of sample proportions,
two time points and fifty percent overlap in sample
(For a sample of size n multiply entries by 2 and divide by n)

$P_{E.1}$	$P_{E.2}$	P_{EE}	$P_{.E2}$	$P_{.E3}$
0.0651	0	0	0	0
0	0.0651	0.0637	0.0358	0
0	0.0637	0.0819	0.0546	0
0	0.0358	0.0546	0.0564	0
0	0	0	0	0.0564

Table 3
Variance of alternative estimation procedures
(For a sample of size n at each period, multiply entries by 2 and divide by n)

Parameter	Procedure		
	Simple	Restricted GLS	Full GLS
$P_{E.}$	0.0326	0.0326	0.0294
P_{EE}	0.0819	0.0397	0.0374
$P_{.E}$	0.0278	0.0258	0.0255
$P_{EE}/P_{.E}$	0.0290	0.0229	0.0220
$P_{.E} - P_{E.}$	0.0429	0.0367	0.0353

were observed at the first time and one half are new to the sample. We take as our vector of observations the vector containing the proportion of elements in category 1 in the one half of the sample that is not observed the second time [denoted by $P_{E.1}$], the proportion of elements in category 1 at time 1 in the remaining half of the sample [denoted by $P_{E.2}$], the proportion of elements that are in category 1 at both time 1 and time 2 for the portion of the sample that is observed at both time periods [denoted by P_{EE}], the proportion of the elements in category 1 at time 2 for the elements that are observed at both times [denoted by $P_{.E2}$], and the proportion of elements in category 1 at time 2 for the portion of the sample that is observed only at time 2 [denoted by $P_{.E3}$].

We assume simple random sampling. Then, because the statistics are sample proportions, it is easy to write down the covariance matrix of the vector of five estimators. A multiple of that covariance matrix is given in Table 2. To obtain the covariance matrix for a sample of size n at each time period, divide every entry in the table by n and multiply by two. In Table 3 we give the variance of alternative estimation procedures. In the first column is the variance of the procedure that uses as the estimator of the first period proportion only the elements appearing in the first period sample. To estimate the fraction appearing in category 1 (employed) both at time 1 and time 2, the simple procedure uses only the overlap elements, and to estimate the number in the first category at time 2, it uses only the sample observed at time 2. Thus, if we have a sample of 200 elements at each time period, the first period sample of 200 elements is used to estimate the first probability. The 100 elements observed at both time 1 and time 2 are used to estimate the proportion of the elements in category 1 at both time 1 and time 2, and the 200 elements observed at time 2 are used to estimate the time 2 proportion.

The last column is the variance of the best linear unbiased estimators constructed using generalized least squares. The estimators are constructed from the vector of five basic statistics and the covariance matrix of that vector. This estimator is of the form

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} Y, \quad (1)$$

where V is given in Table 2, $\beta = (P_{E\cdot}, P_{E\cdot}, P_{EE})$,

$$X' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

and Y is the five-dimensional vector of direct estimates,

$$Y' = (\bar{P}_{E\cdot 1}, \bar{P}_{E\cdot 2}, \bar{P}_{EE}, \bar{P}_{E2}, \bar{P}_{E3}).$$

The second column of Table 3 gives the variance of the restricted least squares estimators, where the restriction is that the estimator for the first period must be the estimator obtained from the initial sample. This would be an appropriate procedure if the agency never made a revision in the once published estimates. For example, the Bureau of Labor Statistics in the United States does not revise the unemployment statistics. Once released, they are the official estimates. Of course, the United States unemployment statistics are based on a more complicated sample and are based on a survey that is conducted over a longer period of time than our example.

To describe the restricted generalized least squares estimator of Table 3, let the model be

$$Y = X\beta + e,$$

where X is a fixed $n \times k$ matrix and

$$E\{ee'\} = V.$$

The generalized least squares estimator of β , with some elements of β restricted to be certain linear combinations of Y can be constructed as follows. Consider the Lagrangian

$$(Y - X\beta)' V^{-1} (Y - X\beta) - 2 \sum_{i=1}^b \lambda_i (\Gamma_i \beta - g_i),$$

where Γ_i is a fixed row vector and b is the number of restrictions. The solution to this minimization problem is defined by

$$\begin{pmatrix} X' V^{-1} X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X' V^{-1} Y \\ g \end{pmatrix},$$

where $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_b)$, $\Gamma' = (\Gamma'_1, \Gamma'_2, \dots, \Gamma'_b)$ and $g' = (g_1, g_2, \dots, g_b)$. If we replace g by the linear combination GY , the equation becomes

$$\begin{pmatrix} X' V^{-1} X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} X' V^{-1} \\ G \end{pmatrix} Y.$$

This equation defines the restricted estimator of β as a linear function of Y . Hence the variance of the estimator of β is the upper $k \times k$ portion of

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} V \left(\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} \right)'.$$

This is not the only way to compute the restricted generalized least squares estimator. An alternative estimator of level and change that leaves the previous estimator unchanged is the composite estimator. See, for example, Wolter (1979).

Several points are illustrated by this small example. First, with a correlation of 0.591 between employment at the two time periods, the improvement in the current estimate of employment from using generalized least squares is modest, about 10%. On the other hand, there is a very large improvement in the variance of the estimate of P_{EE} from using generalized least squares. The variance of the generalized least squares estimator of P_{EE} is about 45% of the variance of the simple estimator. The second important point is that the use of restricted generalized least squares to estimate P_{EE} and P_E produces estimates that are nearly as efficient as full generalized least squares. There is about a one percent loss for the estimate of P_E and about a six percent loss for the estimate of P_{EE} .

3. LONGITUDINAL ESTIMATORS

Recall that our definition of a pure panel survey is one in which the same elements are observed at every time point of data collection. The pure panel survey is possible for observations of certain physical units, such as plots of land. In the case of surveys of human populations, the pure panel must be considered to be a figment of the statistician's imagination. In the real world, a fraction of the respondents from the first time are always unavailable at the second time. Good reviews of procedures for missing data are given by Lepkowski (1989) and Little and Su (1989). Also see Little and Rubin (1987), Kalton (1983) and Madow *et al.* (1983).

We have described the rotating panel survey in which the design calls for some elements to leave the study and some elements to enter the study at every time point at which the study is conducted. In this type of survey we might say that we have planned nonresponse for those elements that are rotated out of the sample. Thus, estimation in the presence of nonresponse and estimation for rotating panel surveys are related problems.

Given that one does not obtain data from every respondent at every point in time of a repeated survey, one is faced with a choice among methods of handling planned and unplanned nonresponse. There are two simple, and common, procedures. If the interest is in following individuals over time, then very often the investigator retains in the study only those individuals that responded every time. A weighting procedure may be used to adjust the data using characteristics of the initial respondents and (or) external auxiliary data. This procedure is often used in special one-time studies of a specific population. In such situations the report on the study is released only after the entire study is completed.

The second common type of estimation procedure is to construct estimates for each time period using the data that are available for that time period. This procedure is often used if the survey is repeated regularly, the results are released after each survey, no revisions are made in the releases, and no longitudinal estimates are produced. One-period-at-a-time estimation has the advantage of being very easy to compute at time t because no information from the previous period is used in calculating the current estimators. It generally gives good estimates (not optimal) of the current value, but rather poor estimates of change.

In fact, one might use both of these procedures in a single survey. The Survey of Income and Program Participation (SIPP) conducted by the U.S. Bureau of the Census is a panel survey with a rotating time-of-interview with a four-month recall period. The Census Bureau provides a set of weights at each time of the survey that can be used to construct estimates for that point in time using all individuals that respond at that time point. They also provide (a) the sample of individuals that responded all eight times for the period 1984-1985 with weights for these individuals, (b) the sample of individuals that responded all four times in 1984 with an appropriate weight and (c) the sample of individuals that responded all four times in 1985 and an appropriate weight.

We outline an estimation procedure for a panel survey with nonresponse where the analysis is conducted at the end of the survey. It is assumed that a reasonable fraction of the units respond at all time points of the survey and that longitudinal analysis is of interest. The computational procedure consists of constructing weights for the units with complete response records. Information from respondents with incomplete records constitutes a form of auxiliary information.

The first step in the analysis is to pick a few variables that are very important to the study. The number of variables that can be used will depend upon the sample size. The covariance structure of the vector of estimates composed of the simple estimates for each of these variables for each type of response pattern for each point in time where the estimate is appropriate, is computed. The covariance structure is a function of the response-nonresponse pattern. There are different definitions of simple estimators. For simple random sampling, simple estimators are simple means. For stratified samples, one might define the original vector to include estimates for each stratum. Alternatively, the simple estimator for a stratified sample might weight the responses in each stratum for nonresponse. The vector Y used in (1) is an example of a vector of simple estimates.

Given the vector of simple estimators and the estimated covariance matrix of the vector, improved estimators for each of the time periods is constructed by generalized least squares. For example, if we had a panel study with three time points, there are seven response patterns. These are XXX , $0XX$, $X0X$, $XX0$, $X00$, $0X0$, $00X$, where X denotes response and 0 denotes nonresponse. If we choose two variables of interest, the vector of simple estimates will contain $12 \times 2 = 24$ estimates because there are 12 group-response times associated with the seven response patterns. In this example, generalized least squares would be used to produce six estimates, the estimates for the two variables for each of the three time periods.

The generalized least square estimators for the selected characteristics become control variables for a next stage of estimation. Using regression weighting methods, weights are constructed for the individuals that responded at all time periods. The weights are constructed so that the generalized least squares estimates for each time period are reproduced by the weighted sample of 100% respondents. That is, the time estimates for the chosen variables are used as controls.

The efficiency of the procedure depends upon the correlation between the chosen control variables and the analysis variable. If a control variable is also the analysis variable, the procedure will be very efficient. The procedure is less than fully efficient for the control variables only because a limited amount of information is used in the generalized least squares procedure.

The strong advantage of the outlined procedure is that it produces a single tabulation data set that can be used to construct internally consistent estimates for all reporting times and for all gross change tables. The disadvantage is that estimates for particular points in time are less than fully efficient.

The variance of the procedure can be computed by analogy to the procedures used for double sampling. Let Y be the characteristic of interest. For simplicity, assume a simple random sample at each time. We write the model to be used in estimation as

$$Y_i = \mu_Y + (X_i - \mu_X)\theta + e_i$$

$$\mu_X = E\{X\},$$

$$e_i \sim \text{Ind}(0, \sigma_e^2).$$

Let $\hat{\mu}_X$ be the generalized least squares estimator of μ_X . Then our estimator for the mean of Y is

$$\hat{\mu}_Y = \bar{y} + (\hat{\mu}_X - \bar{x})\hat{\theta},$$

where $\hat{\theta}$ is the vector of regression coefficients obtained in the regression of Y_i on X_i using the set of complete observations, and (\bar{y}, \bar{x}) is the mean vector for the elements observed at every time period. Let m be the number of complete observations. Then the variance of the estimator is, approximately

$$V\{\hat{\mu}_Y\} = m^{-1}\sigma_e^2 + \theta' V\{\hat{\mu}_X\}\theta,$$

where $V\{\hat{\mu}_X\}$ is the covariance matrix of $\hat{\mu}_X$.

The least squares estimator we have described will perform well in most situations. However, it is possible for the estimator to produce negative estimates for quantities known to be non-negative. This is because the estimator is linear and it is possible for some of the weights to be negative. Procedures have been developed to avoid this problem. See Huang and Fuller (1978).

4. THE U.S. NATIONAL RESOURCE INVENTORY

The Iowa State Statistical Laboratory cooperates with the U.S. Soil Conservation Service on a large survey of land use in the United States. The survey was conducted in 1958, 1967, 1975, 1977, 1982, and 1987. A survey is currently being planned for 1992.

The survey collects data on soil characteristics, land use and land cover, potential for converting land not used for crops to cropland, soil and water erosion, and conservation practices. The data are collected by employees of the Soil Conservation Service. Iowa State University has responsibility for sample design and for estimation.

The sample is a stratified sample of the nonfederal area of 49 states (all except Alaska) and Puerto Rico. The sampling units are areas of land called segments. The segments vary in size from 40 acres to 640 acres. Data are collected for the entire segment on items such as urban land and water area. Detailed data on soil properties and land use are collected at a random sample of points within the segment. Generally, there are three points per segment, but 40-acre segments contain two points and the samples in two states contain one point per segment. Some data, such as total land area and area in roads, are collected on a census basis external to the sample survey.

In 1982, the sample contained about 350,000 segments and nearly one million points. The 1987 sample was composed of about 100,000 segments. The majority of the 1987 sample segments were a subsample of the 1982 segments. However, about 1,500 new segments were selected in areas of rapid urban growth. Data were collected on about 280,000 points in 1987.

Table 4
Illustration of estimation procedure

1982	1987				TOTAL
	Cropland	Other	Urban	Roads	
Cropland	26,243	179	13	6	26,441
Other	771	7,114	6	2	7,893
Urban	0	0	623	0	623
Roads	17	4	0	1,038	1,059
1987 TOTAL	27,031	7,297	642	1,046	36,016

For the first time in 1987, it was decided that longitudinal data analysis would be performed for the period 1982-1987. Also for the first time, it was decided that the data were to be made available to the state Soil Conservation Service staff so that they could perform their own analyses.

In 1987, the field personnel were provided with a preprinted work sheet containing the 1982 information for the segment. They entered the information for 1987 on the basis of field observation and aerial photography. Field personnel were permitted to change the 1982 data if they found it to be incorrect. Edit and checking procedures were applied throughout the processing operation.

The sample was designed to produce reasonable estimates for units called Major Land Resource Areas. These areas are defined on the basis of soil and cover characteristics. There are about 180 Major Land Resource Areas in the study area. Also the acreage estimates for any county were to be consistent with the total acreage of that county. There are about 3,100 counties in the sample. Because the sample must provide consistent acreage estimates for both counties and Major Land Resource Areas, the basic tabulation unit is the portion of a Major Land Resource Area within the county. There are 5,530 of these units, which we called MLRAC's.

The design of the sample is a simple form of a panel survey in that the 1987 sample is nearly a subsample of the 1982 sample. It was decided to use as the control variables from the 1982 study, the 1982 acres of 14 major land uses such as cropland, rangeland, forestland, and urban land. In addition, the external information, such as 1987 area in roads, and the segment information, such as 1987 area in urban land, is auxiliary information similar to that obtained from incomplete observations.

Table 4 is a condensed version of an estimation table for one of the states in the survey. It contains only four uses instead of the 14 actually employed in the estimation. The entries in the right column are the 1982 estimates. The entries in the last row for urban land and roads are from the segment data and the external sources, respectively. The vector of six entries, (the first four entries of the last column, 1987 urban land, and 1987 roads) is a vector of totals corresponding to the vector of estimated means, $\hat{\mu}_X$ of Section 3.

The internal estimates of the table are essentially least squares estimates that satisfy the six control totals. In the actual estimation scheme it was necessary to use imputation methods when, for example, a change is reported in the segment data, but there is no corresponding change in the point data.

The design produced large variances for the directly estimated change in small uses such as urban land, farmsteads, and small water bodies. Therefore, a small area estimation scheme was used to construct estimates of change for the major land resource areas within counties.

We used a computer program for small area estimation developed at Iowa State University. The theory for the small area estimation procedure is described in Fuller (1986). Estimated changes in five small land uses for each of the 5,500 MLRAC's were constructed with the small area program. This procedure is essentially an allocation program in that the sum of the MLRAC estimates is the state estimate. Estimates for the entries in Table 4 (with 14 categories) were constructed for each MLRAC.

In this estimation, the small area MLRAC estimates, the external estimate for roads, and the state marginals for cropland were used as controls. The final step in the estimation procedure was the assignment of weights to the point data such that the weighted point data give the estimates of Table 4 for each MLRAC.

To summarize, the final product of the estimation procedure is a tabulation data set of points that permits estimation of complete two-way tables of 1982-1987 land use for any identifiable area designation. The estimates are consistent with previous estimates for major land use categories for the states and are consistent with data from sources outside of the point sample.

Generally speaking, it is not possible to obtain good variance estimates from the tabulation sample, although segment and stratum identification are given in the data set. Simple variance estimates computed with the point data for principal uses, such as cropland, will be too large because of the control on the larger 1982 sample. Proper variance estimation requires the use of double sampling formulas.

5. MEASUREMENT ERROR

Measurement error can have a very large impact on the analysis of data over time. This impact may be moderate in the case of simple means reported at a sequence of times. However, in gross change estimation and in regression estimation, measurement error can be extremely important.

To illustrate the magnitude of measurement error bias in estimators of gross change, let us return to the simple example of Table 1. If the data were collected by a procedure such as that of the U.S. Census Bureau, the work of Chua and Fuller (1987) demonstrates that the interior cells of the two-way table will be seriously biased. Also see Abowd and Zellner (1985), Poterba and Summers (1985), and Singh and Rao (1990). Under the Chua-Fuller model, the response error at the two points in time is assumed to be independent. Also it is assumed that, at each time,

$$\begin{aligned} P\{\text{response} = E | \text{true} = E\} &= 1 - \alpha + \alpha P_E, \\ P\{\text{response} = U | \text{true} = E\} &= \alpha P_U, \\ P\{\text{response} = U | \text{true} = U\} &= 1 - \alpha + \alpha P_U, \\ P\{\text{response} = E | \text{true} = U\} &= \alpha P_E, \end{aligned}$$

where α is the parameter of the response mechanism. Under this model the expected value for the proportion employed at any point in time is the true proportion. A consistent estimator for P_{EE} under the Chua-Fuller model is

$$\hat{\pi}_{EE} = (1 - \alpha)^{-2} \{ \hat{P}_{EE} - \hat{P}_E \cdot \hat{P}_E [1 - (1 - \alpha)^2] \},$$

where \hat{P}_{EE} , \hat{P}_E , and \hat{P}_E are the direct estimators and α is a parameter of the response mechanism. Also see Battese and Fuller (1973). On the basis of the U.S. reinterview data, a value of $\alpha = 0.10$ is not unreasonable. For our example, we have

Table 5
Mean square error of alternative estimators for a sample of 10,000 at
each time and 50% overlap
(Mean square error of measurement error adjusted GLS = 100)

Parameter	Procedure					
	Ordinary			Measurement Error		
	Simple	Rest. GLS	Full GLS	Simple	Rest. GLS	Full GLS
$P_{E.}$	111	111	100	111	111	100
$P_{.E}$	111	101	100	111	101	100
P_{EE}	1071	967	961	250	106	100

$$\begin{aligned}\pi_{EE} &= (0.90)^{-2}\{0.91 - 0.93(0.94)(0.19)\} \\ &= 0.9184.\end{aligned}$$

The corresponding two-way table of proportions adjusted for response error is

$$\begin{pmatrix} 0.9184 & 0.0116 \\ 0.0216 & 0.0484 \end{pmatrix}.$$

In this example, the bias in the direct estimator of P_{EE} is 0.0084. Chua and Fuller estimate the bias to be about 0.0168 in the three-way table that includes the not-in-the-labor-force category. Table 5 contains a comparison of alternative estimation procedures for P_{EE} . A sample of 10,000 is assumed. The first three procedures are those of Table 3. The last three are the three estimators adjusted for measurement error bias. In the variance calculations, α is assumed to have a standard error of 0.01. The estimators of $P_{E.}$ and $P_{.E}$ are not changed by the adjustment for measurement error bias. In this example, the squared bias in the ordinary estimator of P_{EE} is about nine times the variance of the generalized least squares estimator. Thus, the measurement error bias dominates the mean square error of the estimator of P_{EE} .

These results have serious implications for survey design. To illustrate this, we return to the gross change problem. Assume that our objective is to estimate the probability that a person will remain employed for two periods, P_{EE} . We assume that it is possible to conduct independent reinterviews for each point in time, and that interviews at two points in time are independent. We assume that the only interview procedures permitted are:

- A. Interview and reinterview at one of the times.
- B. Interview at time one and interview at time two.

We assume that the response error is unbiased and that a simple two-class (employed and unemployed) model is appropriate. We also assume that the probabilities of correct response depend only on the current class of the respondent. Let the response probabilities be defined in terms of α and let

$$\gamma = (1 - \alpha)^{-2}.$$

Let θ_{ij} denote the ij -th element of the 2×2 matrix of probabilities observed in the reinterview study. That is, θ_{ij} is the probability that an individual responds i on the first interview and j

Table 6
MSE efficiency of MEM to direct

	Sample size, <i>n</i>			
	500	1,000	5,000	10,000
MSE direct/MSE MEM	0.87	1.13	3.22	5.84

on the reinterview. For this simple model we can obtain explicit expressions for the estimators. We have

$$\hat{\gamma} = (\hat{\theta}_{11} - \hat{\theta}_1^2)^{-1}(\hat{\theta}_1 - \hat{\theta}_1^2)$$

and

$$\hat{P}_{11} = \hat{\gamma}(\bar{P}_{11} - \bar{P}_1.\bar{P}_{.1}) + \bar{P}_1.\bar{P}_{.1}$$

where

$$\theta_1 = \theta_{11} + \theta_{12} = \theta_{11} + \theta_{21},$$

$\hat{\theta}_{ij}$, are the estimates from the reinterview study and \bar{P}_{ij} are the estimates from the interviews conducted at the two time periods.

In constructing the estimator, the reinterview study is used only to estimate the measurement error parameter. In fact, the reinterview study could be used in a generalized least squares procedure to improve the estimates of P_{11} , $P_1.$ and $P_{.1}$. Under the assumption that all interviews are of equal cost, it can be demonstrated that about one fourth of the resources should be used for the reinterview study. The relative efficiency of the measurement error procedure to the direct biased procedure is given in Table 6.

In small samples, the direct procedure has a smaller mean square error because of the smaller variance. Recall that only three fourths of the observations furnish information on $P_{EE} = P_{11}$. However, for samples greater than 750, the squared bias dominates the mean square error of the direct procedure and the consistent measurement error procedure has a smaller mean square error. This small example demonstrates the efficacy of surveys containing a component to estimate the parameters of the measurement process.

6. SUMMARY AND CONCLUSIONS

We have reviewed some topics associated with the analysis of repeated data, without attempting a complete discussion of the topic. We have shown that procedures based upon least squares have the potential to provide large gains in efficiency. Because of size and timing considerations, it is not possible to include all available information in the construction of the least squares estimators. Thus, in practice, the statistician must choose a subset of variables to use in the construction of least squares weights. Estimation for a two-period survey conducted by the U.S. Soil Conservation Service was described.

We illustrated the large biases that measurement error can produce in longitudinal estimates such as gross changes estimates. We showed that measurement error methods exist that can be used to construct consistent estimators. The use of one fourth of the available resources to estimate the variance of the measurement error in order to use measurement error estimation methods can be justified.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 68-3A75-8-12 with the Soil Conservation Service, U.S. Department of Agriculture. I thank Margot Tollefson for the computations.

REFERENCES

- ABOWD, J.M., and ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- BATTESE, G.E., and FULLER, W.A. (1973). An unbiased response model for analysis of categorical data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 202-207.
- BATTESE, G.E., HASABELNABY, N.A., and FULLER, W.A. (1989). Estimation of livestock inventories using several area and multiple frame estimators. *Survey Methodology*, 15, 13-27.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- ECKLER, A.R. (1955). Rotation sampling. *The Annals of Mathematical Statistics*, 26, 664-685.
- FULLER, W.A. (1986). Small area estimation as a measurement error problem. *Proceedings of the Conference on Survey Research Methods in Agriculture*, (Ed. D. Faulkenberry), American Statistical Association and NASS, U.S. Department of Agriculture, Washington, D.C.
- GARCIA, P.A., BATTESE, G.E., and BREWER, W.D. (1975). Longitudinal study of age and cohort influences on dietary patterns. *Journal of Gerontology*, 30, 349-356.
- GRAHAM, J.E. (1973). Composite estimation in two cycle rotation sampling designs. *Communications in Statistics*, 1, 419-431.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics, American Statistical Association*, 242-257.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 300-303.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. University of Michigan, Survey Research Center.
- KASPRZYK, D., DUNCAN, G.J., KALTON, G., and SINGH, M.P. (1989). *Panel Surveys*. New York: John Wiley.

- KASPRZYK, D., and McMILLEN, D.B. (1987). SIPP: Characteristics of the 1984 Panel. *Proceedings of the Section on Social Statistics, American Statistical Association*, 181-186.
- LAZARSFELD, P.F., and FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.
- LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- LITTLE, R.J.A., and SU, H.L. (1989). Item Nonresponse. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley.
- MADOW, W.G., OLKIN, I., NISSELSOHN, H., and RUBIN, D.B. (1983). *Incomplete Data in Sample Surveys*. (Three volumes) New York: Academic Press.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- POTERBA, J.M., and SUMMERS, L.H. (1985). Adjusting the gross change data: Implications for Labor Market Dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 81-95.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SINGH, A.C., and RAO, J.N.K. (1990). Adjustments for classification error in gross flows. Unpublished manuscript, Statistics Canada, Ottawa, Canada.
- SMITH, T.M.F., and HOLT, D. (1989). Some inferential problems in the analysis of surveys over time. Paper presented at the 47th session of the International Statistical Institute, Paris.
- WOLTER, K. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Sample Maintenance Based on Peano Keys

KIRK M. WOLTER and RACHEL M. HARTER¹

ABSTRACT

We discuss frame and sample maintenance issues that arise in recurring surveys. A new system is described that meets four objectives. Through time, it maintains (1) the geographical balance of a sample; (2) the sample size; (3) the unbiased character of estimators; and (4) the lack of distortion in estimated trends. The system is based upon the Peano key, which creates a fractal, space-filling curve. An example of the new system is presented using a national survey of establishments in the United States conducted by the A.C. Nielsen Company.

KEY WORDS: Recurring surveys; Sample maintenance; Changing population units; Peano key.

1. INTRODUCTION

We are concerned with recurring surveys conducted over time and the maintenance they require. Let \mathcal{U}_t denote a survey universe at time t , with $t = 0$ denoting the inception of a new survey. We assume a probability sample of units of \mathcal{U}_0 has been selected, and thus that it is feasible to construct unbiased (or at least consistent) estimators of the population total and other parameters of interest. As time goes by, we assume the universe is surveyed repeatedly at regular intervals of time, in part to track the “level” of the population, and in part to measure its “trends”. A panel or a rotation sampling design is usually employed for this purpose (*e.g.*, see Rao and Graham (1964) and Wolter (1979) and the references cited by those authors). In all such surveys of people or their institutions, which is all we concern ourselves with here, the composition of the universe changes with time as births, deaths, and other changes occur to the status of the units. The survey frame, the sampling design, and the schemes for observing or collecting the survey data must be maintained for such change; otherwise, the sample may become excessively biased and cease to be representative of the universe.

The types of maintenance issues that arise in recurring surveys depend in part on the kind of universe under study, in part on the choice of sampling unit, and in part on the interplay between the sampling unit and the universe elemental units. We shall summarize briefly the issues that arise in four different situations:

- (i) establishment surveys with establishment as the sampling unit;
- (ii) establishment surveys with company or some similar cluster of establishments as the sampling units;
- (iii) surveys of people or households with the address or housing unit as the sampling unit; and
- (iv) surveys of people or households with the household or family as the sampling unit.

In this work, we use the words “establishment” and “company” in a generic sense. An establishment may be a retail store, a manufacturing plant, a school, a hospital, a golf course, or any other similar, single-location entity, while the corresponding company would be the corporate, legal entity that owns the retail store, or the school district, and so on. In some cases, of course, the establishment and company will be synonymous, *e.g.*, a single, independent grocery store.

¹ Kirk M. Wolter and Rachel M. Harter, Statistical Research Department, A.C. Nielsen Company, Nielsen Plaza, Northbrook IL 60062, USA.

For case (i), the main universe dynamics include:

- establishments arising from new construction
- reclassified establishments from some out-of-scope category to an in-scope category
- reclassified establishments from one in-scope category to another in-scope category
- reclassified establishments from an in-scope category to an out-of-scope category
- conversion of a structure from residential use to commercial use
- conversion of a structure from commercial use to residential use
- demolition of an existing establishment
- establishment that moves in and out of vacancy status
- changes in the configuration of an establishment, *e.g.*, division into two or more establishments.

Case (ii) is far more complicated than case (i), principally because sampling units are now clusters of elemental units. All of the issues from case (i) apply to single-establishment companies. For multi-establishment companies, we face the following additional dynamics:

- mergers wherein two companies combine to form a new successor company
- acquisitions wherein one company is acquired by another, with the acquiring company as the sole successor company
- joint ventures wherein two companies collaborate to form a new company that may be a subsidiary to both the parent companies
- divestitures wherein a company spins off a new and independent company
- divestitures where a company sells parts of itself to another acquiring company.

In a sense, case (iii) is very similar to case (i) in respect to the kinds of universe dynamics that may arise:

- housing units arising from new construction
- reclassified housing units from some out-of-scope category to an in-scope category
- reclassified housing units from one in-scope category to another
- reclassified housing units from an in-scope category to an out-of-scope category
- conversions from residential to commercial
- conversions from commercial to residential
- demolition of an existing housing unit
- reconfigurations of existing structures, *e.g.*, reconfigurations of apartments within a small multiunit structure.

Note how closely these issues match those for case (i).

Finally, case (iv) is very similar to case (ii) in terms of the composition and complexity of universe change. Maintenance issues include:

- marriage, wherein a new successor family is created, possibly from whole predecessor families or from part families
- new members move into an existing family, either eliminating another family or part of a family
- divorce, wherein successor families may be created from one predecessor family
- family members move away, either to join another existing family or to establish a new family
- births of family members
- deaths of family members
- a whole family moves, thus requiring tracing and perhaps altering field-work assignments.

To handle the universe dynamics listed above, properly reflecting them in the sample, so that sample representativeness is retained over time, the survey organization must design and adopt an explicit system of maintenance. We define a *sample maintenance system* to be a sampling design and a universe updating methodology, possibly specified in the form of simple rules, that permit the statistician to achieve known, nonzero probabilities of inclusion for each of the elemental units in the population for each time period in the recurring survey, or failing that, to weight the survey data properly so as to achieve unbiased or consistent estimators of the population parameters of interest. From cases (i) through (iv) above, it is clear that a maintenance system must perform at least four functions:

- give new elemental units a known, nonzero probability of selection
- account properly for elemental units that may no longer exist in a substantive sense
- not give elemental units multiple chances of selection into the sample; otherwise, if multiple changes are given, the system must appropriately record this information so that adjustments may be made in the estimation procedures
- appropriately update the universe frame so as to facilitate and control the above activities.

A general and necessary rule of thumb for any sample maintenance system is that the system, or the rules that define the system, must treat symmetrically universe changes both within and outside of the sample. If a proposed maintenance rule violates this rule of thumb, then there is risk of bias in estimators of totals and other universe parameters to be estimated. For example, consider two rules that might be used for case (ii) for sampling new companies created as the result of a divestiture. One possibility is to declare the new companies part of the sample *if* their predecessor companies were part of the sample, and otherwise, if their predecessors were not part of the sample, to subject the new companies to a new round of sampling. This rule is seen to give the new companies multiple probabilities of selection, and thus may result in biased estimation unless appropriate adjustments are made in the estimation procedure. (The adjustments we have in mind are related to the multiplicity rules studied by Monroe Sirken (1970) and others.) A second possibility is to declare the new companies part of the sample *if and only if* their predecessor companies were part of the sample. Because this second rule treats symmetrically the universe changes both within and outside of the sample, it is seen to result in unbiased estimation for the survey parameters of interest.

In designing a sample maintenance system, the statistician must be guided not only by the statistical properties of the resulting estimators, but also by the cost, feasibility, and customer acceptance of alternative rules. Some rules may require additional data collection, thus entailing additional cost that must be planned from the inception of a new recurring survey. Certain applications may actually require that additional data be collected retrospectively. This may be impractical, or at the very least, may entail considerable nonsampling error, thus risking bias. Some rules may well be feasible and cost-effective, yet may not satisfy the requirements of the customers or users of the survey data.

Finally, we note that this problem of maintenance is neither new nor newly recognized; for example, maintenance systems have been in place for years in many of the major recurring surveys at Statistics Canada, the United States Bureau of the Census, and the A.C. Nielsen Company. Nevertheless, there is remarkably little literature on this subject. For brief discussions of some maintenance issues, see Wolter *et al.* (1976) for case (ii), Hanson (1978) for case (iii), and Ernst (1989) for case (iv). Also see the broad comments of Duncan and Kalton (1987) on household surveys and Colledge (1989) on business surveys.

In the balance of this article, we focus on case (i), where the establishment is both the sampling and elemental unit. This is the case we face in our establishments surveys at the A.C. Nielsen Company. Section 2 describes one of our major surveys, the Scantrack survey,

and the specific maintenance issues we face in that survey. We also describe some of the key objectives we had in designing a new maintenance system for this survey.

The new maintenance system is based upon a parameter known in mathematics as the Peano key, which creates a fractal, space-filling curve. The Peano key is defined in Section 3, where we also provide several graphical displays for illustration purposes. We close the article in Section 4 by describing the rules that implement our new maintenance system.

2. THE SCANTRACK SURVEY

The Nielsen companies provide information from several marketing surveys. The media surveys, such as Nielsen Television Index and Nielsen Station Index, are based on samples of either housing units or households. Surveys for the packaged goods industry, including Nielsen Food Index, Nielsen Drug Index, and Nielsen Scantrack United States (NSUS), are based on samples of stores. The Single Source service, which ties together consumer purchasing behavior with household television viewing and retail marketing support, is based on both household and store samples. Although sample maintenance is an important issue to each of these surveys, the present discussion will focus on our Scantrack sample of grocery supermarkets, which is the basis for the NSUS service. The Scantrack sample includes 3,000 supermarkets, stratified by 50 metropolitan markets and a remaining United States stratum. Within a market, the sample is further stratified by major chain organizations. The frame is ordered geographically, and a systematic sample is selection within each stratum to achieve proper socio-economic representation. This sample is also representative of store age, store size, and other factors associated with item sales. Although a geographically ordered systematic sample is exceedingly simple and straightforward, the choice of this sample design is justified based on years of experience, as well as the results of empirical studies in which various sample designs were tested on universe data.

Stores in the Scantrack sample are equipped with electronic scanners at the checkout, which read bar codes on packaged goods. Bar codes are called universal product codes or UPC's. When the item is scanned, the transaction is entered into the store's computer where the UPC is matched with the item's price. Each week, the sample stores provide us with total sales movement and price data for every item that is scanned in the store. Since a supermarket typically carries over 10,000 UPC's, we receive and process over 30 million observations per week.

In addition to scanner data, we obtain data on promotion conditions for the items in each of the sample stores, including whether an item was featured in a newspaper advertisement, store display, or store coupon. If an item was featured, we also know the type of newspaper advertisement used and the location of the display within the store.

NSUS reports include estimated sales totals for individual items and aggregates of items for each market and the total United States. A ratio estimator is used, with all-commodity volume as the auxiliary variable. All-commodity volume, or ACV, refers to total sales of all items in a store, usually on an annual basis. ACV tends to be highly correlated with sales of individual items. In addition, the NSUS reports include estimates of sales and sales rates by promotion condition and estimates of year-to-year sales trends.

Continuous maintenance is necessary for the Scantrack sample because the national supermarket universe of approximately 30,500 stores is not static. In a recent 12-month period, approximately 2,200 new supermarkets opened, and 2,450 existing stores went out of business. Another 170 stores were reclassified during the year. Reclassification can result from any of a number of changes. Some smaller grocery stores enter the Scantrack universe when their ACV's surpass the \$2-million-per-year threshold which defines a supermarket. A store might

change name or location, or be expanded through remodeling. Some stores change to an extended or economy format, such as a superstore, warehouse store, or other nontraditional supermarket. In 1979, about 3,800 extended and economy stores accounted for 17% of total supermarket sales. By 1988, the number of extended and economy stores had grown to over 9,000, and they accounted for almost 50% of all supermarket sales (*Progressive Grocer* 1989). Sometimes, individual stores or entire chains are acquired by another organization affecting stratum definitions.

In addition to universe changes, missing or faulty data situations arise that require substitution of sample stores. Some selected sample stores do not scan, and some that do have incompatible scanning equipment. If a store is consistently unable to provide us with usable data, it must be dropped from the sample. Sometimes a request for a sample change within an organization comes from the chain itself. Occasionally, a retailer simply refuses to cooperate.

The principal objectives of our maintenance system for the Scantrack sample are:

- (1) the sample should maintain geographic balance through time
- (2) the system should maintain the sample size through time
- (3) the sample should adhere to principles of probability sampling so as to avoid bias in estimators of total sales, and
- (4) sample changes should not disturb excessively estimates of year-to-year trends.

Geographic balance is a proxy for socio-economic balance. Because different neighborhoods have different purchasing patterns, geographical balance is important to achieving an efficient sample design (*i.e.*, low sampling variability) over a wide range of products. Furthermore, geographic balance is an important factor in our customers' perception of an appropriate sample.

A sample size decrease would adversely affect the standard errors of the estimators, and a sample size increase would adversely affect our costs. Neither outcome is desirable. Furthermore, contracts with chain organizations specify sample sizes and cooperation payments, and any changes would have to be renegotiated. This too is undesirable.

All applications involving Scantrack data require efficient, unbiased estimators of total sales. Manufacturers and retailers need such data for everyday business decisions, such as how much to produce, how much to ship, how much to keep in inventory, and how to allocate store shelf space.

Clients also require reliable year-to-year trend information for managing their businesses. Trend estimates help manufacturers assess the overall health of their businesses. Both manufacturers and retailers benefit from knowing the longer-term performance of all major brands in all product categories.

We describe the maintenance system that has been developed to meet these objectives in section 4. But first, we describe a new geographic ordering scheme in section 3.

3. PEANO KEYS

The Peano key is a parameter that defines a certain fractal, space-filling curve. It provides a mapping from \mathbb{R}^2 to \mathbb{R}^1 such that points in \mathbb{R}^2 or spatial objects can be arranged in a unique order (Peano order) on a list. In the application we have in mind, the spatial objects are sampling units, and the space \mathbb{R}^2 is represented by earth's geographic coordinate system.

We obtain the Peano key by interleaving bits. See Peano (1908), Laurini (1987) and Saalfeld, Fifield, Broome and Meixler (1988). Let $X = X_k \dots X_3 X_2 X_1$ and $Y = Y_k \dots Y_3 Y_2 Y_1$ represent the longitude and latitude of an arbitrary point in k -digit binary form. Then, the corresponding Peano key is $P = X_k Y_k \dots X_3 Y_3 X_2 Y_2 X_1 Y_1$. Also see figure 1 for an example for the case $k = 4$. Note how simple it is to calculate the value of P .

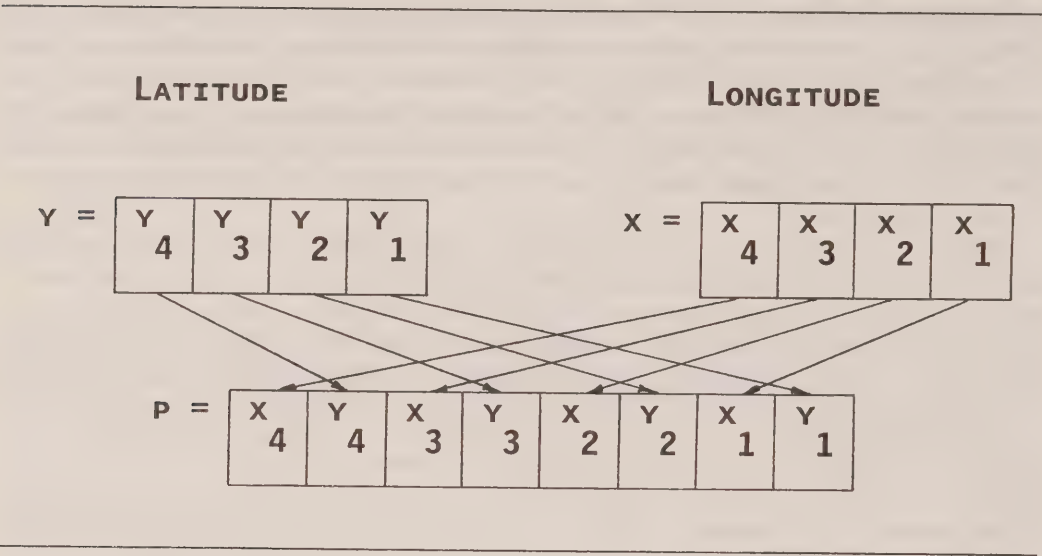


Figure 1. Creating the Peano Key by Bit Interleaving

Given k -digit (for any finite k) latitude longitude coordinates, the spacial “point” represented by the value of P is actually a square in \mathbb{R}^2 . As k increases, the sizes of the squares decrease. In fact, as k tends to infinity, the value of P will tend to represent a specific point in \mathbb{R}^2 .

The space-filling curve created by the values of the Peano key, P , is in the shape of a recursive N . Figure 2 illustrates the N -curve, using a grid of 1024 points. This figure displays the self-similarity feature of fractal images.

The N -curve passes once and only once through each point in space, points being defined as squares whose size is determined by the number of digits carried in the latitude and longitude coordinates. The order of points on the curve (Peano order) is largely preserving of geographic contiguity. Thus, Peano order facilitates proximity searches. Peano order involves a few geographic discontinuities, such as the jump from point 516 to point 517 in figure 2, as does any mapping from \mathbb{R}^2 to \mathbb{R}^1 .

In the specific application we envision here, economic establishments are arranged on a list in Peano order by means of their latitude and longitude coordinates. Probability samples of the establishments may be drawn systematically from the ordered list. Because the earth’s coordinate system is stable, there is no ambiguity in determining the list position of new establishments. Thus, they may be subjected to sampling too.

To illustrate this application, see figure 3 which displays a chain of retail establishments in the United States. Each establishment is described by a double-letter code. This code in natural lexicographic order signifies the Peano order of the establishments.

In the next section, we describe a sample maintenance system that is based upon the establishments’ Peano order.

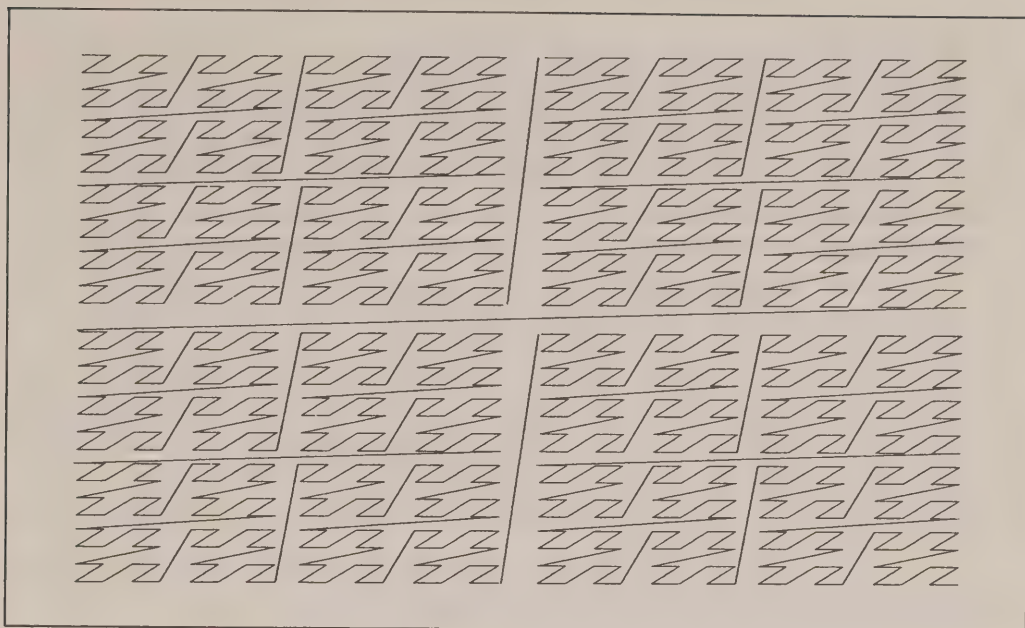
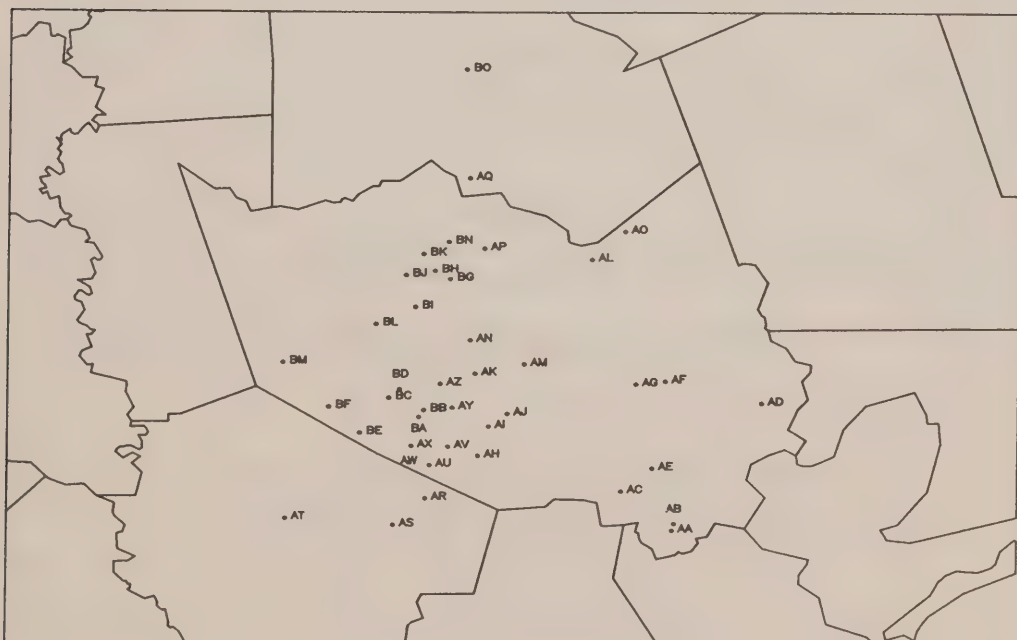


Figure 2. Peano Order Based on 1024 Points



4. RULES FOR MAINTAINING THE SAMPLE

We describe a system for maintaining samples of retail stores, taking proper account of births, deaths, scanning conversions, and other changes in the status of the retail store universe. As stated earlier, we developed the system for applications at the A.C. Nielsen Company.

We consider a given and arbitrary sampling stratum, say of size N , and assume the universe of stores in the stratum is arranged in Peano order. For example, a stratum might include all stores in a given metropolitan market, such as Vancouver or Montreal. Ordering by Peano key values will turn out to be especially well-suited to the maintenance system that follows. Other ordering schemes may be considered for this work so long as they are stable across time and effectively map \mathbb{R}^2 to \mathbb{R}^1 in such fashion as to preserve geographic contiguity and to assign all birth stores a unique position in the ordering.

We assume an original sample is selected systematically with equal probability from the ordered list of stores at time $t = 0$. Let U_{ij} denote the j -th store in the i -th possible systematic sample, for $i = 1, \dots, k$ and $j = 1, \dots, n_i$, where k is the sampling interval and n_i is the size of the i -th possible sample. If $N = nk + r$, $r < k$, then r samples will be of size $n_i = n + 1$ and $k - r$ samples of size $n_i = n$. In what follows, we shall also use the subscript “ i ” to represent the sample actually selected.

Let P_{ij} denote the Peano key value associated with U_{ij} . Let P_L and P_U denote the smallest and largest possible Peano key values within the market under study. Thus,

$$P_L \leq P_{11} < P_{21} < \dots < P_{k1} < P_{12} < \dots < P_{ij} < \dots < P_{kn_k} \leq P_U.$$

Note that we are assuming each store possesses a unique geographic location and thus a unique Peano key value.

Let Y_{tij} denote the value of some characteristic of U_{ij} at time t . A standard unbiased estimator of the population total, Y_t , is

$$\hat{Y}_{ti} = k \sum_{j=1}^{n_i} y_{tij},$$

while the ratio estimator is given by

$$\hat{Y}_{Rti} = \hat{Y}_{ti} X_t / \hat{X}_{ti},$$

where the X -variable is a measure of size and X_t and \hat{X}_{ti} are analogous to Y_t and \hat{Y}_{ti} , respectively.

Define N Peano key segments, S_{ij} , by partitioning the range $[P_L, P_U]$ at the N store values P_{ij} . We let $S_{ij} = [P_{ij}, P_{i+1,j})$, where it will be understood that $P_{k+1,j}$ represents $P_{1,j+1}$. A special definition is needed for the final segment. We define $S_{kn_k} = [P_{kn_k}, P_U] \cup [P_L, P_{11})$ so that the entire Peano range $[P_L, P_U]$ is covered by the N segments. This special definition, which treats the Peano range as if it were on a circle, is needed later to guarantee that all store births are given a nonzero probability of selection. Alternative segmentation schemes may be used without defeating the statistical properties of the maintenance system.

Our maintenance scheme is based upon the Peano key segments. The basic idea is to view the systematic selection process as applying to the segments, with subsampling of stores within the selected segments. Thus, as a formal matter, the segment is the primary sampling unit (PSU), not the store. Of course, as of the time of initial sample selection, there is, by construction, only one store per segment.

4.1 Birth Sampling

At a future point in time, say t' , one or more new stores may open for business. Each new store will be assigned its unique Peano key value, and this value will be an element of one and only one Peano key segment. The Peano key permits us to automatically place new stores in their correct and unique positions on the ordered universe list.

The simplest possible rule for sampling births is the following:

Rule 1. A birth store is selected into the sample if and only if its Peano key value is an element of a selected Peano key segment. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

Given this rule, a birth store is selected with probability $1/k$. This occurs because its segment, which is unique, is selected with probability $1/k$. Unfortunately, Rule 1 does not provide good control of the sample size over time.

To control the sample size, we advocate some form of subsampling within PSU's. Let $U_{ij1}, U_{ij2}, \dots, U_{ijB_{ij}}$ denote the stores in segment S_{ij} . The original store is now labeled U_{ij1} , whereas $U_{ij2}, U_{ij3}, \dots, U_{ijB_{ij}}$ are the birth stores in Peano order. The number, $B_{ij} - 1$, of births in any given segment will be 0, 1, or 2 in most applications. Then we may subsample as described in the following alternative rule.

Rule 1A. A birth store will be subjected to subsampling if and only if its Peano key value is an element of a selected Peano key segment. Associate with $U_{ij1}, U_{ij2}, \dots, U_{ijB_{ij}}$ the probabilities $p_{ij1}, p_{ij2}, \dots, p_{ijB_{ij}}$, where $p_{ijb} > 0$ and $\sum p_{ijb} = 1$. Now choose one of the stores according to this probability measure. Subsampling is independent from one selected segment to the next. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

The probabilities in Rule 1A may be equal or unequal. If unequal, they may be defined in proportion to some preliminary measures of size, or defined so as to accelerate or retard the replacement of the sample.

We observe that our principal maintenance objectives are well-satisfied by Rule 1A. First, the rule maintains geographic balance over time because there is always one unit selected from each of the originally selected segments, which themselves were geographically balanced by virtue of the systematic sampling design. Second, the rule maintains a constant sample size over time because there is always one and only one store selected from each of the originally selected segments. Third, the rule is in accord with strict principles of probability sampling, whereby probabilities of inclusion are known and nonzero, and thus unbiased estimators of population totals are available. Finally, by appropriate choice of the p_{ijb} , we may control distortion in year-to-year trends.

The unconditional probabilities of selection are given by

$$\pi_{ijb} = k^{-1} p_{ijb}$$

for $b = 1, \dots, B_{ij}$. That is, π_{ijb} is equal to the probability of selecting the PSU times the conditional probability of selecting the store, given the selected PSU.

Let $Y_{t'ijb}$ denote the value of the unit U_{ijb} , and let $Y_{t'ij+}$ denote the total for the (i,j) -th PSU. Then, the unbiased estimator of the population total $Y_{t'}$ is given by

$$\hat{Y}_{t'i} = \sum_{j=1}^{n_i} y_{t'ijb} / \pi_{ijb},$$

where $y_{t'ijb}$ is the value of the single unit selected from the (i,j) -th selected segment, with variance

$$\text{Var}\{\hat{Y}_{t'i}\} = \frac{1}{k} \sum_{i=1}^k \left(k \sum_{j=1}^{n_i} Y_{t'ij+} - Y_{t'} \right)^2 + k \sum_{i=1}^k \sum_{j=1}^{n_i} \sigma_{t'ij}^2, \quad (1)$$

where

$$\sigma_{t'ij}^2 = \sum_{b=1}^{B_{ij}} p_{ijb} \left(\frac{Y_{t'ijb}}{p_{ijb}} - Y_{t'ij+} \right)^2.$$

The first term on the right side of (1) is the variance due to the sampling of segments. This is the original variance in the sense that it is the variance expression that applied at the time of original sample selection. The second term on the right side is the variance due to subsampling within segments. Note that $\sigma_{t'ij}^2$ vanishes for any segment in which birth subsampling has not occurred. Note also that the subsampling scheme achieves its minimum variance when, for each given i and j , the probabilities p_{ijb} are defined to be proportional to $Y_{t'ijb}$. In this case, the within component of variance vanishes. For any real application, however, this proportionality condition will be satisfied only approximately.

As usual, a first-order Taylor series approximation may be used to discover the variance of the ratio estimator. See Wolter (1986) for appropriate techniques to estimate the variance of both the unbiased estimator, $\hat{Y}_{t'i}$, and the ratio estimator $\hat{Y}_{Rt'i}$.

As time passes, it will be necessary to periodically update the sample to reflect additional births and other changes in the universe. It may be desirable to schedule the updating at regular intervals of time, so as to facilitate management of the work. We will refer to these intervals as update cycles. Such cycles may occur monthly, bimonthly, quarterly, or at whatever interval makes sense in a particular application. Factors to consider in establishing the frequency of the updating cycles include cost of the updating process; desired accuracy of the estimators of level and trend; and perceptions of the customers or users of the data.

Generally speaking, more frequent updating will cost more, achieve greater accuracy, and be perceived better by customers than less frequent updating.

For an update cycle at any future time t' , Rules 1 or 1A may be used to maintain the sample. New stores are always placed automatically in their correct segment, by their Peano key values, and the subscript b reflects this order at each cycle. To explicitly reflect these ideas, we should have further subscripted the U 's, B 's, p 's, and π 's by time, but we avoided doing so as a notational convenience. The expressions for the estimators of total, $\hat{Y}_{t'i}$ and $\hat{Y}_{Rt'i}$, and their variances remain valid for each t' .

4.2 Updating for Deaths

Rules for maintaining a sample over time must obey an important general principle. They must treat equally both selected and nonselected units. In the case of deaths, this principle implies that all deaths, both those in and out of the sample, must be handled in the same fashion in any sample updating process. If this principle is not followed, the resulting estimators will be biased, and the bias may accumulate over time.

In what follows, we describe procedures for death updating that follow this essential principle. There are two cases to consider: (i) deaths are not known on a universe basis, (ii) deaths are known on a universe basis.

For case (i), we suggest Rule 2.

Rule 2. All deaths in the sample will be known. They should remain in the sample but be set to zero (*i.e.*, $y = 0$) at the time of an update cycle.

This rule permits unbiased estimation of the universe population totals. Deaths cause the estimator variances to increase, and estimators of variance will properly reflect this increase, provided the deaths are retained in the sample with zero values.

For case (ii), we suggest Rule 3.

Rule 3. Remove all deaths from the universe at the time of the next update cycle. Subject only the remaining live cases to sampling, including births.

Rule 3 will cause the store count B_{ij} to change in segments where deaths have occurred, unless births exactly offset deaths. A replacement store will necessarily be selected within a given segment whenever the sample store from the segment has died -- except when there is a death but no birth and $B_{ij} = 0$ -- and a replacement store may be selected even when the sample store is alive and well.

In the exceptional case, where $B_{ij} = 0$, the sample size drops by 1. An interesting problem for future research is to investigate the mean square error of this rule versus that of an alternative rule which selects a replacement store from the same zone of k stores, instead of permitting the sample size to drop by 1. This alternative is conditionally unbiased but unconditionally biased.

Two additional issues must be addressed in handling deaths. The first issue concerns the coordination of birth and death updating. Store births and deaths will occur naturally at irregular intervals, depending upon business conditions and population growth. In some time periods, neither births nor deaths will occur. In other time periods, births may occur but not deaths, or vice versa. While in other periods, both deaths and births will occur. In theory, it would be possible to employ different update cycles for store births and deaths. For example, one might update bimonthly for both births and deaths, but in alternating months. This approach may have advantage in leveling the work load over time. On the other hand, alternating cycles may tend to defeat the ability of the sample to properly measure trends, creating a sawtooth pattern in the store time series as first births are introduced, then deaths dropped, then births, deaths, and so on. On balance, we recommend coincident sample updating for births and deaths so as to preserve trends.

The second issue concerns the handling of deaths during the period from their actual occurrence until the next update cycle. This issue arises only if the frequency of the updating process is less than that of the data-collection process. If the two processes are coincident, then there are no new problems. If updating is the less frequent, then there are two alternatives:

- a) drop the deaths from the sample as soon as they are known to us (to be more precise statistically, this means the deaths are included in the sample with a value of zero)
- b) continue the deaths in the sample by imputing for them until the time of the next update cycle.

Alternative a) is the simplest, cleanest way of proceeding. Aside from the problem of births, it is unbiased and permits correct variance estimators. Because of the birth problem, however, this alternative may have a negative effect on the ability of the sample to properly measure trends. As deaths occur during the first weeks of an update cycle, one can imagine a slight decline in the store time series, not because of fundamental change in economic conditions, but simply because the sample reflects deaths and not births. Alternative b) provides a short term fix to the problem of properly measuring trends. The essential notion here is that by imputing for

deaths, we implicitly make a correction for any births that have occurred since the last update cycle. This fix is not particularly elegant, and it is difficult to frame a rigorous, unassailable technical justification for it. On the other hand, history has shown that populations of economic establishments tend to be stable in the short run. Deaths are often associated with or are compensated by births, with the net size of the population remaining approximately level in the short run. The United States Bureau of the Census has used this alternative in its whole-sale trade survey, with quarterly update cycles and monthly data collection. See Wolter *et al.* (1976).

4.3 Chronically Nonusable Stores or Scanning Conversions

In this final subsection, we present sample maintenance rules for handling stores that are chronically nonusable, such as stores that do not scan; do scan but with such poor discipline as to render their data faulty and nonusable; or refuse to participate in the survey. We shall explicitly discuss nonscanning stores and sample maintenance rules for handling conversions from nonscanning to scanning and vice versa, although the material that follows may be seen to apply more generally to all conditions of chronic nonusability. We shall let A denote the set of scanning stores and B the set of nonscanning stores, where $A \cup B$ spans the entire universe.

First, we treat conversions to scanning. There are two principal cases to consider: (i) scanning status is known for all stores prior to sampling; (ii) scanning status is not known prior to sampling, but is observed after sampling for the selected stores only.

Case (i) is relatively easy to handle. Here is a natural rule:

Rule 4. Do not subject nonscanning stores B to sampling. Sample only from the subuniverse of scanning stores A . As a given nonscanning store converts to scanning, then treat it as a birth, subjecting it to birth sampling. Prior to conversion, non-scanning stores B shall be represented in the universe by utilizing imputation or other missing data techniques.

Given this rule and the prior data (*i.e.*, scanning status) it assumes, the entire survey budget may be allocated to the sample of scanning stores. None of the sample resources need to be committed to nonscanning stores.

To address case (ii), let s denote the selected sample of stores, and let $s_A = s \cap A$ and $s_B = s \cap B$. By assumption, s_A and s_B are not observed until after initial field work is completed. Obviously, all of these sets vary with time, but we suppress explicit time subscripts to simplify the notation.

Sample s_A should be maintained by rules presented elsewhere in this paper for births and deaths. New rules are required to handle s_B . Here is an illustrative rule that treats the stores in s_B as nonrespondents.

Rule 5. At time t , impute for store $U_{ijb} \in s_B$ the value $\hat{y}_{tijb} = x_{tijb} y_{At} / x_{At}$, where x_{tijb} is the value of an auxiliary variable for store U_{ijb} , y_{At} is the sample s_A total for the estimation variable, and x_{At} is the corresponding total for the auxiliary variable. Alternatively, imputation may occur by means of substitution, hot deck/matching, or other means. Now, act as if the data set is complete, applying standard estimators of the survey parameters of interest. At the time U_{ijb} converts to scanning, it shall be deleted from s_B and joined to s_A , and the estimation shall still be performed by means of the standard estimators applied to the completed data set.

Given Rule 5, the effective sample size is reduced because of imputation variance associated with the \hat{y}_{tjib} . Substitution maintains a larger effective sample size than the other rules, but is clearly the most expensive to implement. All rules require limited field work on a continuous basis to monitor the scanning status of $U_{ijb} \in s_B$.

As an alternative to missing data techniques, we may observe the nonscanning stores using an alternative mode of data collection. Depending upon the data to be collected, this could involve a store audit or an interview conducted with store personnel by telephone, mail, or in person. This alternative would likely be more accurate than the imputation-based methods, yet additional cost and time may be involved, as well as burden associated with the management and control of two data collection methodologies.

Finally, we treat conversions of sample stores from scanning to nonscanning. Such conversions are likely to be relatively small in number and are treated here only for completeness. Let $U_{ijb} \in s_A$, i.e., i is a scanning store in the sample. Note that U_{ijb} may be either a store that has scanned since being selected into the sample, or a store that converted to scanning after originally entering the sample as a nonscanner under Rule 5.

Rule 6. At the time U_{ijb} converts to nonscanning, it shall be deleted from s_A , joined to s_B , and subsequently handled by missing data techniques, as in Rule 5. Standard formulae shall be applied to the completed data set. To simplify processing and field work, the method selected shall be identical to the method selected to handle conversions from nonscanning to scanning.

In the bizarre instance in which a store flip-flops repeatedly between scanning and non-scanning, one may handle the store by sequentially applying Rule 5 or 6, as the case may be, each time updating the sets s_A and s_B .

REFERENCES

- COLLEDGE, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97-117.
- ERNST, L. (1989). Weighting Issues for Longitudinal Household and Family Estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.
- HANSON, R.H. (1978). *The Current Population Survey: Design and Methodology, Technical Paper 40*. United States Bureau of the Census. Washington, DC.
- LAURINI, R. (1987). Manipulation of Spatial Objects by a Peano Tuple Algebra, University of Maryland Technical Report CS-TR-1893, College Park, MD.
- PEANO, G. (1908). La Curva di Peano nel Formulario Mathematico. In *Opere Scelte di G. Peano*, 115-116, Vol. I. Edizioni Cremonesi, Roma, 1957.
- PROGRESSIVE GROCER (1989). 56th Annual Report of the Grocery Industry 1989, Vol. 68, No. 4, Part 2, Stamford CT.
- RAO, J.N.K., and GRAHAM, J.R. (1964). Rotation Designs for Sampling on Repeated Occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SAALFELD, A., FIFIELD, S., BROOME, F., and MEIXLER, D. (1988). Area Sampling Strategies and Payoffs using Modern Geographic Information System Technology. Unpublished paper, United States Bureau of the Census, Washington, DC.

- SIRKEN, M. (1970). Household Surveys with Multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- WOLTER, K.M. (1979). Composite Estimation in Finite Population. *Journal of the American Statistical Association*, 74, 604-613.
- WOLTER, K.M. (1986). *Introduction to Variance Estimation*. New York: Springer Verlag.
- WOLTER, K.M. *et al.* (1976). Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, Alexandria, VA.

The Time Series Approach to Estimation for Repeated Surveys

WILLIAM R. BELL and STEVEN C. HILLMER¹

ABSTRACT

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977) suggested the use of signal extraction results from time series analysis to improve estimates in repeated surveys, what we call the time series approach to estimation in repeated surveys. We review the underlying philosophy of this approach, pointing out that it stems from recognition of two sources of variation – time series variation and sampling variation – and that the approach can provide a unifying framework for other problems where the two sources of variation are present. We obtain some theoretical results for the time series approach regarding design consistency of the time series estimators, and uncorrelatedness of the signal and sampling error series. We observe that, from a design-based perspective, the time series approach trades some bias for a reduction in variance and a reduction in average mean squared error relative to classical survey estimators. We briefly discuss modeling to implement the time series approach, and then illustrate the approach by applying it to time series of retail sales of eating places and of drinking places from the U.S. Census Bureau's Retail Trade Survey.

KEY WORDS: Repeated surveys; Time series; Signal extraction; U.S. Retail Trade Survey.

1. INTRODUCTION

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977), hereafter SSJ, suggested the use of signal extraction results from time series analysis to improve estimates in repeated surveys. If the covariance structure of the usual survey estimates (Y_t) and their sampling errors (e_t) for a set of time points is known, these results produce the linear functions of the available Y_t 's that have minimum mean squared error as estimators of the population values being estimated (say θ_t) for θ_t a stochastic time series. To apply these results in practice one estimates a time series model for the observed series Y_t and estimates the covariance structure of e_t over time using knowledge of the survey design.

Section 2 of this paper gives a brief overview of the basic results and framework for the time series approach. Section 3 considers some theoretical issues and section 4 some application considerations for the approach. In section 5 we illustrate the approach with an example using two time series from the Census Bureau's Retail Trade Survey.

2. BASIC IDEAS AND GENERAL CONSIDERATION OF THE TIME SERIES APPROACH

The basic idea in using time series techniques in survey estimation that distinguishes it from the classical approach is the recognition of *two sources of variability*. Classical survey estimation deals with the variability due to sampling – having not observed all the units in the population. Time series analysis deals with variability arising from the fact that a time series is not perfectly predictable (often linearly) from past data. Consider the decomposition:

¹ William R. Bell is Principal Researcher, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A., and Steven C. Hillmer is Professor, School of Business, University of Kansas, Lawrence, KS 66045, U.S.A.

$$Y_t = \theta_t + e_t, \quad (2.1)$$

where Y_t is a survey estimate at time t , θ_t is the population quantity of interest at time t , and e_t is the sampling error. The sampling variability of e_t is the focus of the classical survey sampling approach, which regards the θ_t 's as fixed. From a time series perspective all three of Y_t , θ_t , and e_t can exhibit time series variation, as long as they are random and not perfectly predictable from past data. Standard time series analysis would treat Y_t directly and ignore the sampling error in the decomposition (2.1), not treating e_t explicitly, but only indirectly in the aggregate Y_t . In fact, time series analysts typically behave as if the sampling variation is not present and the true values are actually observed. The most basic thing to keep in mind about the use of time series techniques in survey estimation is that there are two distinct sources of stochastic variation present that are conceptualized, modeled, and estimated differently.

2.1 Signal Extraction Results

Suppose that survey estimates Y_t are available at a set of time points labelled $t = 1, \dots, T$. Let $\underline{Y} = (Y_1, \dots, Y_T)'$ and similarly define $\underline{\theta}$ and \underline{e} so we have $\underline{Y} = \underline{\theta} + \underline{e}$. Assuming the estimates Y_t are unbiased and θ_t and e_t are uncorrelated (see section 3.2)

$$E(\underline{Y}) = E(\underline{\theta}) \equiv \underline{\mu} \equiv (\mu_1, \dots, \mu_T)'$$

$$\Sigma_Y = \Sigma_\theta + \Sigma_e, \quad (2.2)$$

where E denotes expectation over both the sampling and time series model distributions, and Σ_Y is the covariance matrix of \underline{Y} , etc. Here $\underline{\mu}$ and Σ_θ refer to the time series structure of θ_t , which is not subject to sampling variation. If Y_t , θ_t , and e_t do not require differencing, it is well known that, since $\text{Cov}(\underline{\theta}, \underline{Y}) = \Sigma_\theta$, using (2.2) the minimum mean squared error linear predictor of $\underline{\theta}$ can be written

$$\hat{\underline{\theta}} = \underline{\mu} + \Sigma_\theta \Sigma_Y^{-1} (\underline{Y} - \underline{\mu}) \quad (2.3)$$

$$= \underline{\mu} + (I - \Sigma_e \Sigma_Y^{-1}) (\underline{Y} - \underline{\mu}) \quad (2.4)$$

$$= \underline{\mu} + (I + \Sigma_e \Sigma_\theta^{-1})^{-1} (\underline{Y} - \underline{\mu}). \quad (2.5)$$

Another standard result is that the variance of the error of this estimate is

$$\text{Var}(\hat{\underline{\theta}} - \underline{\theta}) = \Sigma_\theta - \Sigma_\theta \Sigma_Y^{-1} \Sigma_\theta = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e. \quad (2.6)$$

If normality of $(\underline{\theta}, \underline{Y})$ is assumed (2.3) – (2.5) give $E(\underline{\theta} | \underline{Y})$, the conditional expectation of $\underline{\theta}$ given \underline{Y} , and (2.6) gives $\text{Var}(\underline{\theta} | \underline{Y})$, the conditional variance.

If Y_t requires differencing the preceding results need to be modified. Assume e_t does not require differencing, but θ_t and Y_t need to be differenced once (*i.e.* by applying $1 - B$ where $BY_t = Y_{t-1}$). Let the differenced data be $W_t = (1 - B)Y_t = (1 - B)\theta_t + (1 - B)e_t$ for $t = 2, \dots, T$. Let $\Delta = [\Delta_{ij}]$ be the $(T - 1) \times T$ differencing matrix with $\Delta_{ii} = -1$, $\Delta_{i,i+1} = 1$, and all other elements zero, and write $\Delta \underline{Y} \equiv \underline{W} = \Delta \underline{\theta} + \Delta \underline{e}$. Then we use

$$\hat{\underline{\theta}} = \underline{Y} - \hat{\underline{e}} = \underline{Y} - \underline{\Sigma}_e \Delta' \underline{\Sigma}_{\bar{w}}^{-1} \Delta (\underline{Y} - \underline{\mu}), \quad (2.7)$$

$$\text{Var}(\hat{\underline{\theta}} - \underline{\theta}) = \underline{\Sigma}_e - \underline{\Sigma}_e \Delta' \underline{\Sigma}_{\bar{w}}^{-1} \Delta \underline{\Sigma}_e. \quad (2.8)$$

The expressions (2.7) and (2.8) also apply when θ_t and Y_t require a more general differencing operator (e.g. seasonal differencing), with appropriate definition of the differencing matrix Δ , as long as e_t does not require differencing. These results are analogous to (2.4) and (2.6), but with $\Delta' \underline{\Sigma}_{\bar{w}}^{-1} \Delta$ playing the role of $\underline{\Sigma}_Y^{-1}$. The results are given in Bell and Hillmer (1990), where their optimality properties are discussed. They were essentially given by Jones (1980), but without real justification.

Scott and Smith (1974) and SSJ used classical signal extraction results equivalent to (2.3) – (2.6) based on covariance generating functions rather than covariance matrices. Bell (1984) considers such results for models involving differencing. Another approach (Binder and Dick 1989, Bell and Hillmer 1989) involves putting time series models for θ_t and e_t in state space form and using the Kalman filter and smoother, which can be viewed as an efficient way to compute the matrix results given above. Also, see Tam (1987) for use of the Kalman filter in an explicitly model-based approach to analysis in repeated surveys. In subsequent discussions we generally refer to the results (2.3) – (2.6), though our remarks easily extend to cover the use of (2.7) – (2.8).

In many cases, for time series Y_t and θ_t that are always positive, we will want to take logarithms of Y_t to help induce stationarity of θ_t and the sampling errors. In such cases we rewrite (2.1) as

$$Y_t = \theta_t (1 + \tilde{u}_t) = \theta_t u_t, \quad (2.9)$$

where $\tilde{u}_t = e_t/\theta_t$ and $u_t = 1 + \tilde{u}_t$. Taking logs we get

$$\log(Y_t) = \log(\theta_t) + \log(1 + \tilde{u}_t) = \log(\theta_t) + \log(u_t). \quad (2.10)$$

Letting $\underline{\mu}$ and $\underline{\Sigma}_\theta$ now refer to $\log(\underline{\theta}) \equiv (\log(\theta_1), \dots, \log(\theta_T))'$, and $\underline{\Sigma}_Y = \underline{\Sigma}_\theta + \underline{\Sigma}_u$ refer to $\log(\underline{Y})$, analogous to (2.4) our estimate is

$$\log(\hat{\underline{\theta}}) = \underline{\mu} + [I - \underline{\Sigma}_u \underline{\Sigma}_Y^{-1}] (\log(\underline{Y}) - \underline{\mu}). \quad (2.11)$$

The analogues to (2.6) – (2.8) are obvious. To estimate $\hat{\theta}_t$ we use $\exp[\log(\hat{\theta}_t)]$; alternatively, one could use $\exp[\log(\hat{\theta}_t) + \text{Var}(\log(\hat{\theta}_t) - \log(\theta_t))/2]$ for a more “unbiased” estimate of θ_t with minimum mean squared error (see Granger and Newbold 1976).

Notice that (2.3) – (2.6) require knowledge of $\underline{\mu}$ and any two of $\underline{\Sigma}_Y$, $\underline{\Sigma}_\theta$, and $\underline{\Sigma}_e$ (the third can be obtained from (2.2)). In practice these will not be known exactly and will need to be estimated. Thus, the true minimum mean squared error linear predictor $\hat{\underline{\theta}}$ cannot be obtained exactly and (2.6) or (2.8) understates the mean squared error (MSE) since it does not account for modeling errors. (See Binder and Dick (1989) and Eltinge and Fuller (1989).) The basic assumption underlying the application of the preceding results, which we shall call the time series approach to survey estimation, is that $\underline{\mu}$ and $\underline{\Sigma}_Y$ can be well-estimated from the time

series data on Y_t through a time series model, and Σ_e can be well-estimated using survey microdata and knowledge of the survey design (possibly also using a model). We discuss these issues further in section 4 and illustrate the approach with the example of section 5.

2.2 Some General Considerations of the Time Series Approach

Smith (1978), Jones (1980), and Binder and Dick (1986) review and discuss the approach known as Minimum Variance Linear Unbiased Estimation (MVLU). While both the MVLU and time series approaches can use data from time points other than t in estimating θ_t , they differ in that MVLU regards the θ_t 's as fixed and still only treats one source of variation, that due to sampling. MVLU was developed for cases (such as many rotating panel surveys) where more than one direct estimate of θ_t is available for each t and the e_t 's are correlated over time due to overlap in the survey design. The use of Y_j for $j \neq t$ in estimating θ_t then comes from generalized least squares results and the correlation of the e_t 's. We can see the distinction in terms of our results for the simple case (2.1) where only one direct estimate, Y_t , of θ_t is available, by letting $\text{Var}(\theta_t) \rightarrow \infty$ to get the MVLU. Then $\Sigma_\theta^{-1} \rightarrow 0$ and (2.5) becomes $\hat{\theta} = \bar{Y}$, so without multiple estimates of θ_t the MVLU just uses Y_t to estimate θ_t . These remarks apply generally to composite estimation (Rao and Graham 1964, Wolter 1979), which is often used as an approximation to MVLU.

One question that may arise regarding the time series approach is why one should consider θ_t a stochastic time series? This issue has been discussed by SSJ and at length by Smith (1978). They observe that (1) users of data from repeated surveys treat the data Y_t as a stochastic time series in modeling and would do the same with θ_t if it were available (as it essentially is for surveys with very low levels of error), and (2) classical results (e.g. Patterson 1950) for estimation in repeated surveys (MVLU) assume a time series structure for the individual units in the population, while maintaining the anomalous position that θ_t , which is a function of these individual units (such as the total), is a sequence of fixed, unrelated quantities. In fact, if we assume θ_t is a sequence of fixed, unrelated quantities, then data through any time point are irrelevant to the future behavior of the true series θ_t . If this were the case, then there would be little point in doing the survey in the first place. The data would be out of date as soon as they were published. The real questions here are whether or not we can estimate the time series structure of θ_t and e_t well enough to make beneficial use of this in survey estimation, how worthwhile these benefits may be, and what risks are involved in doing so?

Along with opportunities for improving estimation in repeated surveys, the time series approach offers potential for improved results in other problems where typically only one of the two sources of variability is recognized. It also can potentially unify these as subproblems under one general approach. Such problems include preliminary estimation in repeated surveys (Rao, Srinath, and Quenneville 1989); seasonal adjustment (Wolter and Monsour 1981, Hausman and Watson 1985, Pfeffermann 1991); time series trend estimation and the related problem of detection of statistically significant change over time (Smith 1978); benchmarking, the reconciling of results from a repeated survey with the results from another survey or census estimating the same population characteristics (Hillmer and Trabelsi 1987, Trabelsi and Hillmer 1990); and inference about time series properties of the true series θ_t relevant to economic models (Bell and Wilcox 1990).

Finally, we note that the decomposition (2.1) or (2.10) does not allow for nonsampling errors, nor does the time series approach treat them explicitly. Whether nonsampling error is generally more or less of a problem for the time series approach than for the classical approach is unclear, but one may wish to consider the possible effects of known or suspected nonsampling errors on the time series estimators when applying them in particular situations.

3. THEORETICAL CONSIDERATIONS

We now obtain some theoretical results relevant to the time series approach, and some properties of the resulting estimators.

3.1 Consistency of Time Series Estimators

Following Fuller and Isaki (1981) we let Y_t^ℓ (from the ℓ^{th} sample at time t) be a sequence of estimators of the characteristic θ_t^ℓ of the ℓ^{th} population at time t where the populations and samples for $\ell = 1, 2, \dots$ are nested. (See their paper for details.) Define $\underline{Y}^\ell, \underline{\theta}^\ell, \underline{\varepsilon}^\ell, \underline{\mu}^\ell, \underline{\Sigma}_Y^\ell, \underline{\Sigma}_\theta^\ell, \underline{\Sigma}_\varepsilon^\ell, \hat{\theta}_t^\ell$, and $\hat{\theta}_t^\ell$ in the obvious fashion. We consider what happens to the time series estimators $\hat{\theta}^\ell$ when the estimators \underline{Y}^ℓ are consistent, i.e. $Y_t^\ell \rightarrow \theta_t^\ell$ in some fashion as $\ell \rightarrow \infty$ for $t = 1, \dots, T$, with T , the length of the series, remaining fixed. For now we assume $\underline{\mu}^\ell, \underline{\Sigma}_\theta^\ell$, and $\underline{\Sigma}_\varepsilon^\ell$ are known for each ℓ , which generally means the time series models (including their parameter values) for the components are known. Since $\underline{\mu}^\ell$ and $\underline{\Sigma}_\theta^\ell$ are really superpopulation parameters for the time series, θ_t^ℓ , we wish to estimate, we shall assume these are the same for each population ℓ , that is, $\underline{\mu}^\ell \equiv \underline{\mu}$ and $\underline{\Sigma}_\theta^\ell \equiv \underline{\Sigma}_\theta$ (a positive definite matrix) for all ℓ . This is also partly for convenience since we could get the same results assuming $\underline{\mu}^\ell \rightarrow \underline{\mu}$ and $\underline{\Sigma}_\theta^\ell \rightarrow \underline{\Sigma}_\theta$ as $\ell \rightarrow \infty$.

From (2.5) it would appear that $\underline{Y}^\ell \rightarrow \underline{\theta}^\ell$ would imply $\hat{\theta}^\ell \rightarrow \underline{\theta}^\ell$ as long as $\underline{\Sigma}_\varepsilon^\ell \rightarrow 0$. This condition suggests we need mean square convergence of Y_t^ℓ to θ_t^ℓ . We thus consider estimators Y_t^ℓ of θ_t^ℓ such that $E[(Y_t^\ell - \theta_t^\ell)^2] = E[(e_t^\ell)^2] \rightarrow 0$ as $\ell \rightarrow \infty$. Since $E[(e_t^\ell)^2] = \text{Var}(e_t^\ell) + [E(e_t^\ell)]^2$ this implies both $\text{Var}(e_t^\ell) \rightarrow 0$ and $E(e_t^\ell) \rightarrow 0$. Assuming $Y_t^\ell \rightarrow \theta_t^\ell$ in mean square for $t = 1, \dots, T$ thus implies $\underline{\Sigma}_\varepsilon^\ell \rightarrow 0$. We can now establish

Result 3.1: Consider $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_T)'$ given by (2.4). If $Y_t^\ell \rightarrow \theta_t^\ell$ in mean square as $\ell \rightarrow \infty$ for $t = 1, \dots, T$, then $\hat{\theta}_t^\ell \rightarrow \theta_t^\ell$ in mean square as $\ell \rightarrow \infty$ for $t = 1, \dots, T$.

Proof: From $\underline{Y}^\ell = \underline{\theta}^\ell + \underline{\varepsilon}^\ell$ with $\underline{\Sigma}_\varepsilon^\ell \rightarrow 0$ we have $\underline{\Sigma}_Y^\ell \rightarrow \underline{\Sigma}_\theta$ (even if $\underline{\theta}^\ell$ and $\underline{\varepsilon}^\ell$ are correlated.) From (2.4) we have

$$\hat{\theta}^\ell - \underline{\theta}^\ell = (\underline{Y}^\ell - \underline{\theta}^\ell) - \underline{\Sigma}_\varepsilon^\ell (\underline{\Sigma}_Y^\ell)^{-1} (\underline{Y}^\ell - \underline{\mu}). \quad (3.1)$$

The first term on the right converges to 0 in mean square; the second has mean 0 and variance $\underline{\Sigma}_\varepsilon^\ell (\underline{\Sigma}_Y^\ell)^{-1} \underline{\Sigma}_\varepsilon^\ell \rightarrow 0$ as $\ell \rightarrow \infty$. Since both terms converge to 0 in mean square so does $\hat{\theta}^\ell - \underline{\theta}^\ell$.

Convergence in probability is a more familiar concept in survey sampling. If $Y_t^\ell \rightarrow \theta_t^\ell$ as $\ell \rightarrow \infty$ in probability for $t = 1, \dots, T$ this does not guarantee $\underline{\Sigma}_\varepsilon^\ell \rightarrow 0$, which is mean square convergence, a stronger condition. If we assume there are random variables ζ_t with finite variance such that $|e_t^\ell| \leq \zeta_t$ (almost surely) uniformly in ℓ , then $Y_t^\ell \rightarrow \theta_t^\ell$ in probability implies $Y_t^\ell \rightarrow \theta_t^\ell$ in mean square (Chung 1968, p. 64). Therefore, using Result 3.1, we have

Result 3.2: If $Y_t^\ell \rightarrow \theta_t^\ell$ in probability as $\ell \rightarrow \infty$ for $t = 1, \dots, T$ and there exist random variables ζ_t with finite variance such that $|Y_t^\ell - \theta_t^\ell| \leq \zeta_t$ (almost surely) uniformly in ℓ , then $\hat{\theta}_t^\ell \rightarrow \theta_t^\ell$ in probability as $\ell \rightarrow \infty$ for $t = 1, \dots, T$.

These consistency results show that if the errors in the original estimates Y_t of θ_t are small ($\underline{\Sigma}_\varepsilon$ is small) then the errors $\hat{\theta}_t - \theta_t$ will be small as well. From (3.1) we see this is because $\hat{\theta} - \underline{Y}$ becomes small as $\underline{\Sigma}_\varepsilon$ becomes small, thus when there is little error in the original estimates Y_t the time series approach will not change them much. Binder and Dick (1986) have noted this phenomenon, and also pointed out that in this case it does not matter what time series model is used. That is, the convergence to 0 of (3.1) depends only on $\underline{\Sigma}_\varepsilon^\ell \rightarrow 0$ and not on $\underline{\mu}$ or $\underline{\Sigma}_\theta$. Thus, the consistency results extend to allowing $\underline{\mu}, \underline{\Sigma}_\theta$, and also $\underline{\Sigma}_\varepsilon^\ell$ to be replaced by estimates $\hat{\underline{\mu}}^\ell, \hat{\underline{\Sigma}}_\theta^\ell$, and $\hat{\underline{\Sigma}}_\varepsilon^\ell$ (which will generally come from estimated models – see sections 4 and 5), as long as $\hat{\underline{\mu}}^\ell$ and $\hat{\underline{\Sigma}}_\theta^\ell$ converge to something as $\ell \rightarrow \infty$ (it doesn't matter

what as long as the limit of $\hat{\Sigma}_\theta^\ell$ is positive definite) and $\hat{\Sigma}_e^\ell \rightarrow 0$, which should generally hold when $\Sigma_e^\ell \rightarrow 0$. It is also obvious that these results extend to the nonstationary case where $\hat{\theta}$ is given by (2.7) instead of (2.4). While the results show that the time series estimates behave sensibly in the situation of small error in the original estimates Y_t , the gains from the time series approach will come in the opposite case – when $\text{Var}(e_t)$ is large.

We can extend the consistency results to the case where we take logarithms and estimate $\log(\theta_t)$ using (2.11). In this case let $\Sigma_u^\ell = \text{Var}(\log(u_t^\ell))$ where $\log(u_t^\ell) \equiv (\log(u_{t1}^\ell), \dots, \log(u_{tT}^\ell))'$. If we are taking logarithms it is reasonable to assume Y_t^ℓ and θ_t^ℓ remain bounded away from 0, say $|Y_t^\ell| \geq \kappa$ and $|\theta_t^\ell| \geq \kappa$ (almost surely) for all t and ℓ for some constant $\kappa > 0$.

Result 3.3: If $Y_t^\ell \rightarrow \theta_t^\ell$ in mean square as $\ell \rightarrow \infty$ for $t = 1, \dots, T$, then $\log(Y_t^\ell) \rightarrow \log(\theta_t^\ell)$ and $\log(\hat{\theta}_t^\ell) \rightarrow \log(\theta_t^\ell)$ in mean square as $\ell \rightarrow \infty$ for $t = 1, \dots, T$.

Proof: The analogue to (3.1) is

$$\log(\hat{\theta}^\ell) - \log(\theta^\ell) = (\log(Y^\ell) - \log(\theta^\ell)) - \Sigma_u^\ell (\Sigma_Y^\ell)^{-1} (\log(Y^\ell) - \mu).$$

If we can show $\Sigma_u^\ell \rightarrow 0$ we will have the result since this implies $\log(Y^\ell) \rightarrow \log(\theta^\ell)$ in mean square, and the second term on the right behaves exactly as that in (3.1). Notice

$$E[(\tilde{u}_t^\ell)^2] = E[(e_t^\ell)^2 / (\theta_t^\ell)^2] \leq (E(e_t^\ell)^2) / \kappa^2 \rightarrow 0 \quad \text{as } \ell \rightarrow \infty,$$

thus $E[(\tilde{u}_t^\ell)^2] = E[(u_t^\ell - 1)^2] \rightarrow 0$. This implies $\text{Var}(u_t^\ell) \rightarrow 0$ and $E(u_t^\ell) \rightarrow 1$. By Jensen's inequality (Chung 1968, p. 45), since $\exp(\cdot)$ is a convex function,

$$1 \leq \exp(E[\log(u_t^\ell)^2]) \leq E(\exp[\log(u_t^\ell)^2]) = E[(u_t^\ell)^2].$$

But $E[(u_t^\ell)^2] = \text{Var}(u_t^\ell) + [E(u_t^\ell)]^2 \rightarrow 1$ so $\exp(E[\log(u_t^\ell)^2]) \rightarrow 1$ implying $E[\log(u_t^\ell)^2] \rightarrow 0$. This yields $\text{Var}(\log(u_t^\ell)) \rightarrow 0$ as desired.

As before we could get a convergence in probability result by imposing a boundedness condition on the $\log(u_t^\ell)$. Having $\log(\hat{\theta}_t)$ as an estimate of $\log(\theta_t)$, we have the following Corollary to Result 3.3 for using $\exp[\log(\hat{\theta}_t)]$ as an estimate of θ_t .

Corollary 3.4: If $Y_t^\ell \rightarrow \theta_t^\ell$ in mean square as $\ell \rightarrow \infty$ for $t = 1, \dots, T$, then (see (2.11)) $\exp[\log(\hat{\theta}_t^\ell)] \rightarrow \theta_t^\ell$ in probability as $\ell \rightarrow \infty$ for $t = 1, \dots, T$.

Proof: Since $\log(\hat{\theta}_t^\ell) \rightarrow \log(\theta_t^\ell)$ in mean square implies convergence in probability, the result follows since $\exp(\cdot)$ is a continuous function (Chung 1968, p. 66).

An analogous result obviously holds for using $\exp[\log(\hat{\theta}_t^\ell) + \text{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell))/2]$ to estimate θ_t , since then $\text{Var}(\log(\hat{\theta}_t^\ell) - \log(\theta_t^\ell)) \rightarrow 0$ as $\ell \rightarrow \infty$.

3.2 Uncorrelatedness of θ and e

Standard time series signal extraction results corresponding to (2.3) – (2.8) typically assume and θ_t and e_t are uncorrelated with each other at all leads and lags (equivalent to independence under normality). Previous papers on the time series approach to repeated survey estimation have merely assumed this, but since θ_t and e_t depend on the same population units it is not obvious that this assumption is valid. Fortunately, we can establish that it is valid under fairly general conditions. (Tam (1987) discusses how this fails under an explicitly model-based approach.)

We let y_{it} be the value of the characteristic of interest for the i^{th} unit in the population at time t , and let $\Omega_t = \{y_{it} : i = 1, \dots, N_t\}$ be the collection of all N_t of these units. We consider time points $t = 1, \dots, T$ and let $\underline{\Omega} = (\Omega_1, \dots, \Omega_T)'$. The y_{it} are random variables, as is $\theta_t = \theta_t(\Omega_t)$, which is a function of the y_{it} . The sample at time t , s_t (denoting the indices, not the values, of the units selected), has probability of selection $p(s_t | \underline{\Omega})$. The estimator Y_t of θ_t is a function of the values y_{it} for the units sampled, thus a function of both Ω_t and s_t , i.e. $Y_t = Y_t(\Omega_t, s_t)$. We could let Y_t depend on the sample at times other than t , but we ignore that here for simplicity.

We consider estimators Y_t of θ_t that are *design unbiased*, which we shall define as $E(Y_t | \underline{\Omega}) \equiv \sum_{s_t} Y_t p(s_t | \underline{\Omega}) = \theta_t$. We could alternatively define design unbiasedness as $E(Y_t | \Omega_t) \equiv \sum_{s_t} Y_t p(s_t | \Omega_t) = \theta_t$, and then would need to assume the sample selection process is such that $p(s_t | \underline{\Omega}) = p(s_t | \Omega_t)$, so $E(Y_t | \underline{\Omega}) = E(Y_t | \Omega_t)$. If the sample design is noninformative then s_t and $\underline{\Omega}$ are independent, implying $p(s_t | \underline{\Omega}) = p(s_t | \Omega_t) = p(s_t)$, and either definition of design unbiasedness reduces to $\sum_{s_t} Y_t p(s_t) = \theta_t$. This is the usual definition, which generally assumes the y_{it} , and so Ω_t and θ_t , are fixed. (The assumption $p(s_t | \underline{\Omega}) = p(s_t | \Omega_t)$ allows the sample selection process at time t ($p(s_t | \underline{\Omega})$) to depend on the population values at time t (Ω_t), but assumes the population values at time points other than t (Ω_j for $j \neq t$) offer no additional information on s_t beyond that in Ω_t . This might occur if sampling was with probability proportional to the size of an auxiliary variable at time t that was correlated with the y_{it} only at time t .) The assumptions we make here might even be generalized.

Result 3.5: If Y_t is design unbiased for all t then θ_t and e_t are uncorrelated time series.

Proof: Consider $\text{Cov}(\theta_t, e_j)$ for any two time points t and j . Since Y_j is design unbiased $E(e_j | \underline{\Omega}) = E(Y_j - \theta_j | \underline{\Omega}) = 0$, implying $E[E(e_j | \underline{\Omega})] = E(e_j) = 0$. Also $E(\theta_t \cdot e_j | \underline{\Omega}) = \theta_t \cdot E(e_j | \underline{\Omega}) = 0$ implying $E(\theta_t \cdot e_j) = 0$. Thus $\text{Cov}(\theta_t, e_j) = E(\theta_t \cdot e_j) - E(\theta_t)E(e_j) = 0$.

Comment: If $E(e_j | \underline{\Omega})$ does not depend on $\underline{\Omega}$ then e_j is said to be "mean independent" of $\underline{\Omega}$, which is known to be a stronger condition than e_j and $\underline{\Omega}$ uncorrelated, though not as strong as stochastic independence (unless we have normality). This shows that actually we only need $E(e_t | \underline{\Omega}) = E(Y_t | \underline{\Omega}) - \theta_t$ to not depend on $\underline{\Omega}$ for θ_t and e_t to be uncorrelated time series. This would cover cases where Y_t has a constant additive bias (not dependent on Ω_t) as an estimate of θ_t , or, using approximate Result 3.6 which follows, a constant percentage (multiplicative) bias.

We now consider the logarithmic decomposition (2.10) when the Y_t are design unbiased. We assume that \tilde{u}_j is $O_p(r_\ell)$ where $r_\ell \rightarrow 0$ as $\ell \rightarrow \infty$ in the superpopulation framework of the previous section, omitting the superscript ℓ from random variables here for convenience. (See Wolter (1985, p. 222) for definition of the order in probability notation $O_p(r_\ell)$. For example, when estimating a population mean we would often have $\text{Var}(\tilde{u}_j) \leq K/n_{j\ell}$ where K is some constant and $n_{j\ell}$ is the sample size at time j in the ℓ^{th} population. Then $\tilde{u}_j = O_p(n_{j\ell}^{-.5})$ from Wolter (1985, theorem 6.2.1).) From a Taylor series linearization of $\log(u_j) = \log(1 + \tilde{u}_j)$ we have from Wolter (1985, theorem 6.2.2)

$$\log(u_j) = \tilde{u}_j + O_p(r_\ell^2). \quad (3.2)$$

Using this we obtain the following.

Result 3.6: If Y_t is design unbiased for all t and \tilde{u}_j is $O_p(r_\ell)$, then to terms that are $O_p(r_\ell^3)$, $\log(\theta_t)$ and $\log(u_t)$ are uncorrelated time series.

Proof: From theorem 6.2.5 of Wolter (1985) $\text{Cov}(\log(\theta_t), \log(u_j)) = \text{Cov}(\log(\theta_t), \tilde{u}_j) + O_p(r_t^3)$. Notice $E(\tilde{u}_j | \Omega) = E(e_j | \Omega)/\theta_j = 0$ implies $E(\tilde{u}_j) = 0$, and $E(\log(\theta_t) \tilde{u}_j | \Omega) = \log(\theta_t)E(\tilde{u}_j | \Omega) = 0$ implies $E(\log(\theta_t) \tilde{u}_j) = 0$, so $\text{Cov}(\log(\theta_t), \tilde{u}_j) = 0$, establishing the result.

3.3 Design-Based Properties of Signal Extraction Estimates

Unconditionally, $\hat{\theta}$ in (2.3) is unbiased ($E(\hat{\theta}) = E(\theta) = \mu$) and has minimum MSE given by (2.6). It is easy to see that this is not the case when viewed from a design-based perspective. Suppose we begin with design-unbiased estimators Y , i.e. $E(Y | \Omega) = \theta$. From (2.2) and (2.4) we have $\hat{\theta} - \theta = (I - \Sigma_e \Sigma_Y^{-1}) \varepsilon - \Sigma_e \Sigma_Y^{-1}(\theta - \mu)$. With some algebra, we can show the design bias, variance, and MSE of $\hat{\theta}$ are given by

$$\begin{aligned} E(\hat{\theta} | \Omega) - \theta &= -\Sigma_e \Sigma_Y^{-1}(\theta - \mu), \\ \text{Var}(\hat{\theta} - \theta | \Omega) &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_\theta \Sigma_Y^{-1} \Sigma_e, \\ E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)' | \Omega] &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e \\ &\quad - \Sigma_e \Sigma_Y^{-1} [\Sigma_\theta - (\theta - \mu)(\theta - \mu)'] \Sigma_Y^{-1} \Sigma_e. \end{aligned} \quad (3.3)$$

From a design-based perspective we see use of $\hat{\theta}$ trades bias for a reduction in variance, since $\Sigma_e - \text{Var}(\hat{\theta} - \theta | \Omega)$ is a positive semidefinite matrix. Whether this reduces the conditional MSE (3.3) below Σ_e , the MSE of Y , depends on the last two terms in (3.3), and in turn on θ . There can be particular realizations of θ for which the conditional MSE of $\hat{\theta}$ exceeds Σ_e , though on average signal extraction reduces the MSE by $\Sigma_e \Sigma_Y^{-1} \Sigma_e$, since the unconditional expectation of the bracketed term in (3.3) is zero. (Of course, (3.3) is unusable in practice since it depends on θ .) Also, as noted earlier, modeling error will contribute additional MSE to $\hat{\theta}$, so another fundamental question, more difficult to answer (see Eltinge and Fuller 1989), is how the real unconditional MSE of $\hat{\theta}$ compares to Σ_e ?

4. APPLICATION CONSIDERATIONS

Application of the time series approach to survey estimation requires estimation of the autocovariance structure of the sampling errors, estimation of the mean and autocovariance structure of the signal, and computation of the estimates $\hat{\theta}_t$ and $\text{Var}(\hat{\theta}_t - \theta_t)$ as discussed in section 2. The first two generally involve use of time series models, and are discussed in some detail in Bell and Hillmer (1989). Here we make some general remarks. We assume the Y_t are design unbiased estimators of the θ_t . We illustrate application of the methods in the next section with two time series from the Census Bureau's Retail Trade Survey.

Sampling error autocovariances, $\text{Cov}(e_t, e_{t+k})$, can be estimated in an analogous fashion to sampling variances, $\text{Var}(e_t)$, which is done routinely and for which many methods are available. (See Wolter 1985.) In practice, there may be difficulties in linking survey microdata over time to directly estimate sampling error covariances. Nevertheless, in what follows we assume we have available such estimates $\widehat{\text{Cov}}(e_t, e_{t+k})$ for some set of time points t and lags k . Unfortunately, if there is a substantial amount of sampling error present (the situation where time series methods can make a difference), such autocovariance estimates are likely to have high variances themselves. This suggests some sort of averaging to improve the autocovariance estimates.

First, if we assume e_t is covariance stationary, so $\text{Cov}(e_t, e_{t+k}) \equiv \gamma_e(k)$ depends on k but not t , then each $\widehat{\text{Cov}}(e_t, e_{t+k})$ is estimating $\gamma_e(k)$ and we can simply average them, *i.e.* take $\hat{\gamma}_e(k) = (T - k)^{-1} \sum_t \widehat{\text{Cov}}(e_t, e_{t+k})$ if we have $\widehat{\text{Cov}}(e_t, e_{t+k})$ for $t = 1, \dots, T - k$. Alternatively, $\widehat{\text{Corr}}(e_t, e_{t+k}) = \widehat{\text{Cov}}(e_t, e_{t+k}) / [\widehat{\text{Var}}(e_t) \widehat{\text{Var}}(e_{t+k})]^{.5}$ can be averaged over t to estimate $\text{Corr}(e_t, e_{t+k})$, which also depends on k but not t , and the variance estimates can be averaged as before.

Now suppose we are assuming e_t is relative covariance stationary, so $\text{Cov}(e_t/\theta_t, e_{t+k}/\theta_{t+k}) = \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) \equiv \gamma_u(k)$ depends on k but not t . If \tilde{u}_t is $O_p(r_t)$ for all t , as in section 3.2, then from (3.2) and theorem 6.2.5 of Wolter (1985), $\text{Cov}(\log(u_t), \log(u_{t+k})) = \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) + O_p(r_t^3) \approx \gamma_u(k)$. Taking $\widehat{\text{Cov}}(e_t, e_{t+k}) / (Y_t Y_{t+k})$ as estimates of $\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k})$, these can be averaged over t to estimate $\gamma_u(k)$. Alternatively, using corollary 5.1.5 of Fuller (1976) we can show that $\text{Corr}(\log(u_t), \log(u_{t+k})) = \text{Corr}(\tilde{u}_t, \tilde{u}_{t+k}) + O_p(r_t^3)$, and taking as estimates of $\rho_u(k) \equiv \text{Corr}(\tilde{u}_t, \tilde{u}_{t+k})$, $\{\widehat{\text{Cov}}(e_t, e_{t+k}) / Y_t Y_{t+k}\} / \{[\widehat{\text{Var}}(e_t) \widehat{\text{Var}}(e_{t+k})]^{.5} / Y_t Y_{t+k}\} = \widehat{\text{Corr}}(e_t, e_{t+k})$, we can average the estimated autocorrelations of e_t over t to estimate $\rho_u(k)$, which are approximately the autocorrelations of $\log(u_t)$. Relative variance estimates can be averaged as before.

Actually, the usual survey estimates of variances and autocovariances will be estimating $\text{Var}(e_t | \underline{\Omega})$ and $\text{Cov}(e_t, e_{t+k} | \underline{\Omega})$. These estimates may also be suitable as estimates of $\text{Var}(e_t)$ and $\text{Cov}(e_t, e_{t+k})$, *e.g.* if they make sense from a model-based perspective. If not, and if Y_t is design unbiased so $E(e_t | \underline{\Omega}) = 0$, then averaging autocovariance estimates over time still makes sense. First, if e_t is assumed stationary, then $\gamma_e(k) \equiv \text{Cov}(e_t, e_{t+k}) = E[\text{Cov}(e_t, e_{t+k} | \underline{\Omega})]$, so we can average estimates of $\text{Cov}(e_t, e_{t+k} | \underline{\Omega})$ to estimate $\gamma_e(k)$. Or if e_t is relative covariance stationary, then since $E(\tilde{u}_t | \underline{\Omega}) = E(e_t | \underline{\Omega})/\theta_t = 0$, $\gamma_u(k) \equiv \text{Cov}(\tilde{u}_t, \tilde{u}_{t+k}) = E[\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k} | \underline{\Omega})] = \text{Cov}(\log(u_t), \log(u_{t+k})) + O_p(r_t^3)$, and estimates of $\text{Cov}(\tilde{u}_t, \tilde{u}_{t+k} | \underline{\Omega})$ can be averaged to estimate $\gamma_u(k)$. It is less clear how to justify averaging estimates of conditional (on $\underline{\Omega}$) correlations, since $E[\text{Corr}(e_t, e_{t+k} | \underline{\Omega})] \neq \text{Corr}(e_t, e_{t+k})$, though this may be true to a sufficient approximation. In general, approaches to estimation of sampling error autocovariance structures bear more investigation.

Given an estimate of the sampling error covariance structure, and using any relevant information about the design of the survey, we can attempt to determine a time series model and its parameters to closely reproduce this structure. This is illustrated in the example of section 5.

We now turn to developing a model for the signal, θ_t . Since the behavior of most published time series Y_t is dominated by their signals (otherwise, they would not be published), in developing models for signals θ_t we can draw on experience modeling time series Y_t without allowing for sampling error. Such experience suggests use of nonlinear transformations, differencing, and regression mean functions in the model for θ_t will be important. The logarithm is the most common nonlinear transformation used in time series, and taking $\log(Y_t)$ lets us model $\log(\theta_t)$ through (2.10), with consequences for the sampling error discussed above. The following remarks are given in terms of use of (2.1), but apply equally well to use of (2.10). While other transformations could be considered, they would not generally yield a convenient decomposition of transformed Y_t in terms of transformed θ_t and some sampling error. Choosing between taking logarithms or not transforming seems sufficient for modeling many series.

Assuming e_t has mean zero (implied by design unbiasedness) and does not require differencing, θ_t and Y_t will require the same differencing and have the same mean function. The mean function can often be modeled with a linear regression function, $\mu_t = \underline{X}_t' \underline{\beta}$, for some vector of regression variables \underline{X}_t and parameters $\underline{\beta}$. We often use ARIMA

(autoregressive-integrated-moving average) models to account for the needed differencing and to explain the autocovariance structure of the differenced θ_t . A convenient approach to developing the θ_t model is to first model Y_t ignoring the sampling error, and then use a model with the same regression terms and ARIMA order for θ_t . The parameters of the θ_t model can then be estimated using the time series data for Y_t and the previously developed model for e_t , holding the parameters in the model for e_t fixed. Diagnostic checking may suggest modifications to the θ_t model. The final fitted model can then be used in the signal extraction estimation of θ_t . The model fitting and signal extraction computations are not trivial; Kalman filter/smoothing algorithms are discussed in Bell and Hillmer (1989). These have been implemented in some software recently developed in cooperation with members of the time series staff of the Statistical Research Division of the Census Bureau. This software was used in the analysis of the next section.

5. EXAMPLE: U.S. RETAIL TRADE SURVEY – SALES OF EATING AND DRINKING PLACES

As an illustrative example we analyze time series of sales (in millions of dollars) of Eating Places and of Drinking Places, which are estimated in the monthly U.S. Retail Trade Survey. The Retail Trade Survey has a list panel of large businesses that are selected into the sample with certainty and report sales every month, and 3 rotating list panels of smaller businesses that are selected into the sample by stratified simple random sampling. There is also a rotating panel area sample covering companies not in the list universe. Quarterly, a sample of new firm births is introduced, and firm deaths as determined from activity checks are removed from the sample. The rotating panels report current month and previous month sales at intervals of 3 months for the list sample and 6 or 12 months for the area sample. Horvitz-Thompson (HT) estimates of current and previous months' sales are constructed; the resulting time series shall be denoted Y'_t and Y'_{t-1} . From these composite estimators are constructed as described in Wolter (1979). The final composite estimates will make up our time series Y_t . (While it might be interesting to instead analyze Y'_t and Y'_{t-1} directly, these estimates are not saved for a long enough period of time for seasonal time series modeling.) Sampling variances are estimated using the random group method (Wolter 1985, chapter 2) for the list sample with 16 random groups, and the collapsed stratum method for the area sample. Further information on the survey is given in Isaki *et al.* (1976), Wolter *et al.* (1976), Wolter (1979), Garrett, Detlefsen and Veum (1987), Bell and Wilcox (1990).

There are several complicating factors in the survey. The sample is redesigned and independently redrawn about every five years, with new samples having been introduced in September of 1977, and January of 1982 and 1987. This produces a break in the covariance structure of e_t every five years, which can be handled by the Kalman filter/smoothing as discussed in Bell and Hillmer (1989). We shall use data from September, 1977 through December, 1986, so there is one redrawing of the sample near the middle of our series. When a new sample is introduced approximate MVLU estimates are used for the first three months before switching to the composite estimates (Wolter 1979). This introduces a transient effect into the sampling error autocorrelations that we shall ignore. Finally, the monthly estimates are benchmarked to annual totals estimated from an annual survey and from the economic census taken every five years. To avoid this complication we use data that are not benchmarked. The reader should be aware, however, that for this reason the data used here do not agree with published estimates.

Table 1
Sampling Error Correlations for Horvitz-Thompson Estimates

	Lag					
	4	8	12	16	20	24
Eating Places Averaged ¹	.72	.71	.79	.63	.65	.77
From (5.1) ²	.75	.69	.81	.60	.53	.61
Drinking Places Averaged ¹	.70	.67	.78	.60	.60	.61
From (5.1) ²	.72	.66	.80	.56	.50	.59
Number of Correlations Averaged	23	19	15	11	7	3
Weights Used in Determining $\hat{\phi}$'s	1	1	1	.5	0	0

¹ Raw estimates of $\text{Corr}(e'_t, e'_j)$ and $\text{Corr}(e'_{t-1}, e'_{j-1})$ were available for all pairs of months from January, 1973 through March, 1975. Averages of the correlations for the lags shown were taken after applying Fisher's transformation, and the results then transformed back.

² Correlations are shown from model (5.1) for $m = 4$ with parameters $\hat{\phi}^4 = .604$, $\hat{\phi}_{12} = .723$ (Eating Places) and $\hat{\phi}^4 = .580$, $\hat{\phi}_{12} = .714$ (Drinking Places). These parameter values were determined to minimize the weighted sum of squared deviations of the correlations from model (5.1) and the averaged correlations using the weights shown. Lags 20 and 24 were not used (given zero weight) because of the small number of correlation estimates available at these lags.

5.1 Development of Sampling Error Models

Our first step will be to develop a model for the correlation structure of the sampling errors. Let us write $Y'_t = \theta_t + e'_t$ for the current month (t) HT estimate, and $Y'_{t-1} = \theta_{t-1} + e'_{t-1}$ for the previous month ($t - 1$) HT estimate. We shall use the same models for e'_t and e'_{t-1} . Estimates of $\text{Corr}(e'_t, e'_{t-1})$ are extremely high – typically .98 or higher. While this is partly artificial (due to businesses reporting the same figure for current and previous month sales, and possibly due to the way missing values are imputed), in the absence of other information it is difficult to distinguish characteristics of e'_t from those of e'_{t-1} .

Since the three rotating panels in the survey are drawn (approximately) independently (Wolter 1979), auto- and cross-correlations for (e'_t, e'_{t-1}) should be nonzero only for lags that are multiples of 3. Estimates of such lag correlations can be averaged over time assuming correlation stationarity. While estimates of lag correlations are not regularly produced for the Retail Trade Survey, this was done as part of a special study using micro-data (random group totals) from the Retail Trade Survey sample for January, 1973 through March, 1975, albeit at a time when the survey had four rotating list panels. Lacking more recent data, we “averaged” the correlations at lags 4, 8, 12, 16, 20, and 24 for e'_t and e'_{t-1} . (This was actually done after applying Fisher's transformation $.5 \log((1 + r)/(1 - r))$, to make the distribution of the transformed correlations more symmetric, and then transforming the results back.) The results are shown in Table 1. They show fairly strong positive correlation in the sampling errors, and evidence of seasonality from the correlations at lag 12. A possible model given such data is

$$(1 - \phi^m B^m)(1 - \Phi B^{12})e'_t = v_{1t}, \tag{5.1}$$

where $m = 4$ for the 4-panel survey, with the same model assumed for e'_{t-1} with $v_{2,t-1}$ replacing v_{1t} . (v_{1t} and $v_{2,t-1}$ are white noise with variance σ_v^2 .)

A particularly convenient property of (5.1) is that if the sampling error in each panel would follow (5.1) with $m = 1$ if it were observed every month, then for any number m (that is a divisor of 12) of independent panels reporting successively, e'_t follows (5.1). This allows us to use the 4-panel survey results in Table 1 to estimate ϕ^4 and Φ , and (assuming $\phi > 0$) convert these to estimates of ϕ^3 and Φ , the parameters of the model for the current 3-panel survey. This was done by finding ϕ^4 and Φ to minimize the sum of squared deviations of the correlations from (5.1) with those of Table 1. (Lags 20 and 24 were dropped, and lag 16 given a weight of .5, due to the smaller number of correlation estimates that were averaged together at these higher lags.) This resulted in $\hat{\phi}^3 = .685$, $\hat{\Phi} = .723$ for Eating Places, and $\hat{\phi}^3 = .664$, $\hat{\Phi} = .714$ for Drinking Places. The resulting correlations for $m = 4$ from (5.1) are shown in Table 1, and may be compared to the averaged correlations. More formal statistical estimation procedures for ϕ^3 and Φ , as well as a possible test of model fit, could be considered. (We may pursue this later if sampling error autocorrelation estimates can be produced from more recent micro-data from the 3-panel survey.)

We make the further assumption that $\text{Corr}(e'_t, e'_{t-1-k}) = \rho \text{Corr}(e'_t, e'_{t-k})$ for all k . To justify this, note the population regression of e'_{t-1-k} on e'_{t-k} is $\rho e'_{t-k} + \epsilon$, where if ϵ is not uncorrelated with e'_t , at least it is certainly small since $\text{Var}(\epsilon) = (1 - \rho^2)\text{Var}(e'_t)$ and ρ is very near 1. With this assumption (5.1) leads to the following bivariate model for (e'_t, e'_{t-1}) :

$$(1 - \phi^3 B^3)(1 - \Phi B^{12}) \begin{bmatrix} e'_t \\ e'_{t-1} \end{bmatrix} = \begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} \quad \text{Var} \begin{bmatrix} v_{1t} \\ v_{2,t-1} \end{bmatrix} = \sigma_v^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad (5.2)$$

with $\rho = \text{Corr}(v_{1t}, v_{2,t-1}) = \text{Corr}(e'_t, e'_{t-1})$. Estimates of $\text{Corr}(e'_t, e'_{t-1})$ are regularly produced and were available for 1982 through 1986. Averaging these (with Fisher's transformation) produced $\hat{\rho} = .985$ for Eating Places and $\hat{\rho} = .986$ for Drinking Places.

We can now use (5.2) to derive a model for the sampling error of the linear form of the composite estimator (Wolter 1979), which is given by

$$\begin{aligned} Y'_t{}'' &= (1 - \beta)Y'_t + \beta(Y'_{t-1}{}' + Y'_t - Y'_{t-1}) \quad (\text{preliminary estimator}), \\ Y'_{t-1} &= (1 - \alpha)Y'_{t-1}{}' + \alpha Y'_{t-1}{}' \quad (\text{final estimator}). \end{aligned} \quad (5.3)$$

In the (3-panel) retail trade survey, values of $\alpha = .8$, $\beta = .75$ are used. It is easily seen that (5.3) also holds for the sampling errors, *i.e.* with Y replaced by e . We can use the resulting relations to derive the following equation for e_t in terms of e'_t and e'_{t-1} :

$$(1 - .75B)e_t = .2e'_t{}'' - .75e'_{t-1}{}' + .8e'_t. \quad (5.4)$$

Using (5.2) and (5.4) we then get

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})e_t = .2v_{2t} - .75v_{2,t-1} + .8v_{1t}. \quad (5.5)$$

The right hand side is a first order moving average process (Box and Jenkins 1976, p. 121) whose parameters can be determined given estimates of σ_v^2 and ρ . Thus, (5.5) would yield an ARMA model for e_t .

Rather than pursue this further, we shall instead make the rather strong assumption that a model of the same form holds for $\log(u_t)$ in $\log(Y_t) = \log(\theta_t) + \log(u_t)$, thus

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})\log(u_t) = (1 - \eta B)c_t. \quad (5.6)$$

Table 2
Coefficients of Variation (CV)¹ for Retail Sales Estimates

	Horvitz-Thompson	Final Composite ²	Signal Extraction ³	
	CV	CV	Low	High
Eating Places	.042	.025	.017	.023
Drinking Places	.088	.052	.032	.038

¹ CV = (Relative Variance)⁻⁵.
² The values for the final composite estimator are obtained using models (5.7a,b).
³ The values for signal extraction actually vary over time, being highest at the end of the series and lowest near the middle. We show the lowest and highest values, which are attained for both series in January 1982 (low) and December 1986 (high). The signal extraction variances are not symmetric in time because the sample redraw in January 1982 is not exactly at the center of the series.

We do this because estimates of sampling variance for these series are highly dependent on the level of the series; estimates of relative variance are much more stable over time. We also assume we can use estimates of relative variance and of ρ in determining η and σ_c^2 . Estimates Y'_t , Y'_{t-1} , $\widehat{\text{Var}}(e'_t)$ and $\widehat{\text{Var}}(e'_{t-1})$ were available for 1982 through 1986. The resulting relative variance estimates were used in the spirit of maximum likelihood estimation for the lognormal distribution – taking the average of the logs of the relative variance estimates, adding one half of the sample variance of the logged estimates to this, and exponentiating the results. (Merely averaging the relative variance estimates produced similar results.) This was done separately for $\text{Rel Var}(Y'_t)$ and $\text{Rel Var}(Y'_{t-1})$, and these two results were then averaged, producing a common relative variance estimate that is constant over time. The results are shown in Table 2 under the heading “Horvitz-Thompson”. Using these and the $\hat{\rho}$ ’s given earlier, one can solve for η and σ_c^2 for the right side of (5.6). The resulting sampling error models are

$$(1 - .75B)(1 - .685B^3)(1 - .723B^{12}) \log(u_t) = (1 + .130B)c_t \tag{5.7a}$$

(Eating Places) $\hat{\sigma}_c^2 = 1.948 \times 10^{-5}$

$$(1 - .75B)(1 - .664B^3)(1 - .714B^{12})\log(u_t) = (1 + .134B)c_t \tag{5.7b}$$

(Drinking Places) $\hat{\sigma}_c^2 = 9.301 \times 10^{-5}$.

One can use the method of McLeod (1975,1977) to solve for $\text{Var}(\log(u_t))$ in these models, which is an estimate of the relative variance of the final composite estimator. The results are shown in Table 2. The corresponding coefficients of variation, .025 for Eating Places and .052 for Drinking Places, are quite close to estimates published in the Census Bureau’s Monthly Retail Trade Reports that are obtained more directly.

5.2 Time Series Modeling and Signal Extraction

Figures 1a,b show plots of the time series of final composite estimates Y_t for Eating Places and for Drinking Places, respectively. To develop models for θ_t we shall begin by modeling the Y_t series directly. Both series show trends and strong seasonality, with the magnitude of the seasonal fluctuations larger the higher the level of the series. This suggests taking logarithms and the need for differencing; both are typical for economic time series. Examination

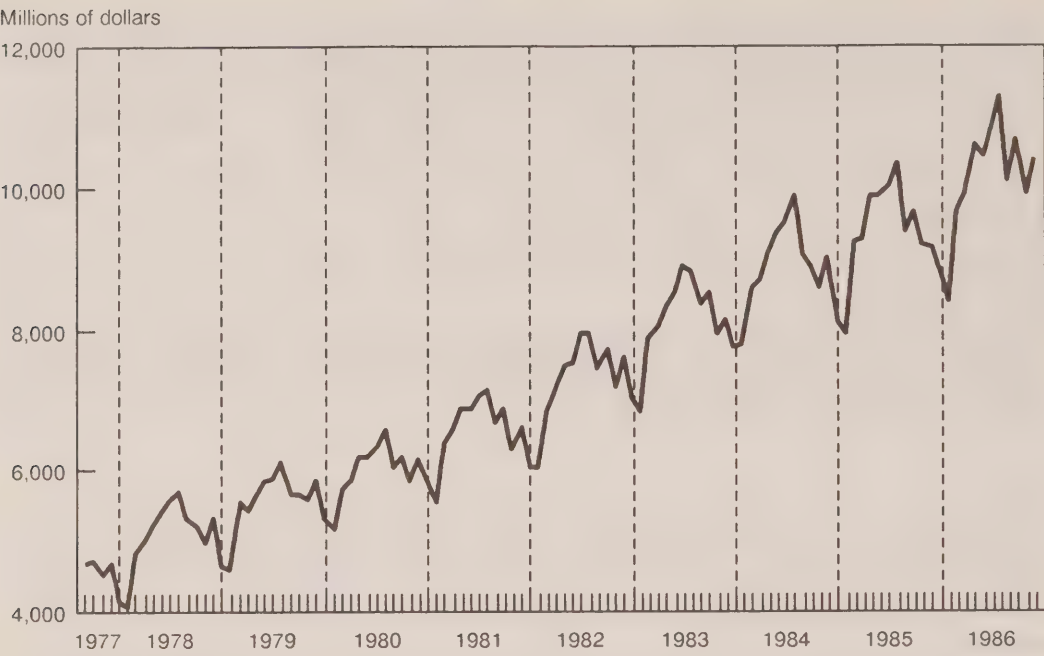


Figure 1.a Retail Sales of Eating Places – Composite Estimates (not benchmarked)

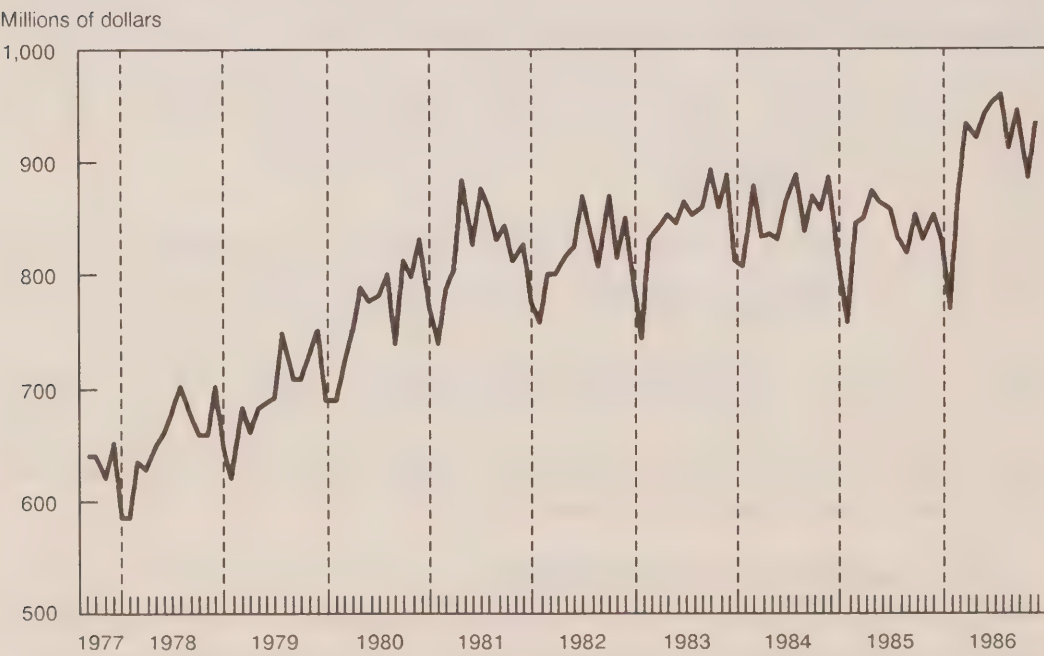


Figure 1.b Retail Sales of Drinking Places – Composite Estimates (not benchmarked)

of sample autocorrelations for $\log(Y_t)$ and its differences suggested the difference operator $(1 - B)(1 - B^{12})$ for both series. Retail trade series are known to contain trading-day variation, which can be modeled by including seven regression variables in the model: X_{1t} = number of Mondays in month t , ..., X_{7t} = number of Sundays in month t . Following Bell and Hillmer (1983), a more convenient parameterization is obtained by using instead the variables $T_{1t} = X_{1t} - X_{7t}$ (number of Mondays - number of Sundays), ..., $T_{6t} = X_{6t} - X_{7t}$ (number of Saturdays - number of Sundays), $T_{7t} = \sum_1^7 X_{it}$ (length of month t). To identify the ARMA structures, the autocorrelations and partial autocorrelations of the residuals from regressions of $(1 - B)(1 - B^{12}) \log(Y_t)$ on $(1 - B)(1 - B^{12})T_{it}$, $i = 1, \dots, 7$, were examined. This suggested an ARIMA (0,1,2)(0,1,1)₁₂ model for Eating Places, and an ARIMA (0,1,3)(0,1,1)₁₂ model for Drinking Places. The resulting estimated models were

$$(1 - B)(1 - B^{12}) \left[\log(Y_t) - \sum_i \beta_i T_{it} \right] = (1 - .25B - .22B^2)(1 - .79B^{12}) a_t$$

(Eating Places) $\hat{\sigma}_a^2 = .000230$ (5.8a)

$$(1 - B)(1 - B^{12}) \left[\log(Y_t) - \sum_i \beta_i T_{it} \right] = (1 - .21B - .15B^2 + .03B^3)(1 - .56B^{12}) a_t$$

(Drinking Places) $\hat{\sigma}_a^2 = .000587$ (5.8b)

For brevity, we omit the estimates of the trading-day parameters. While the lag 2 and lag 3 moving average parameters in (5.8b) are small, we shall retain them since we shall only use (5.8a,b) as starting points for modeling $\log(\theta_t)$ for both series.

Taking models of the form of (5.8a,b) for $\log(\theta_t)$ with models (5.7a,b) for $\log(u_t)$, the parameters of the models for $\log(\theta_t)$ were estimated. For both series the seasonal moving average parameters were estimated to be very near 1(.985 for Eating Places and .992 for Drinking Places), implying nearly deterministic seasonality that can be modeled by cancelling a $(1 - B^{12})$ from both sides of the θ_t model and instead including a trend constant and a seasonal regression function of the form $\sum_1^{11} \gamma_i M_{it}$, where M_{1t} is 1 in January, -1 in December, and 0 otherwise, ..., M_{11t} is 1 in November, -1 in December, and 0 otherwise (Bell 1987). Estimation of the resulting models produced the following:

$$(1 - B) \left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it} \right] = .00762 + (1 - .20B - .29B^2)b_t$$

(Eating Places) $\hat{\sigma}_b^2 = .000139$ (5.9a)

$$(1 - B) \left[\log(\theta_t) - \sum_i \hat{\beta}_i T_{it} - \sum_i \hat{\gamma}_i M_{it} \right] = .00352 + (1 - .18B - .09B^2 - .42B^3)b_t$$

(Drinking Places) $\hat{\sigma}_b^2 = .000244$ (5.9b)

We again omit the estimates of the regression parameters. We do not provide standard errors for the ARMA parameters; doing so for models of the sort used here is a topic for further research, made particularly difficult here by the unrealistic assumption that the sampling error

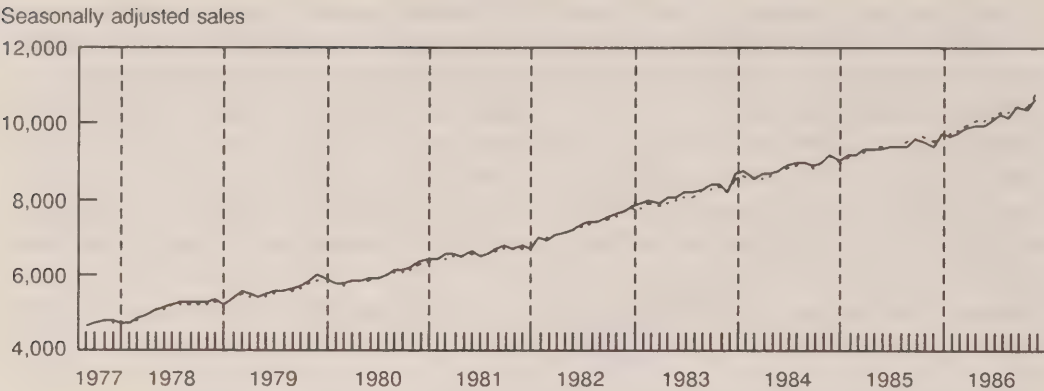


Figure 2.a Eating Places: Composite (solid) and Signal Extraction (dotted) Estimates

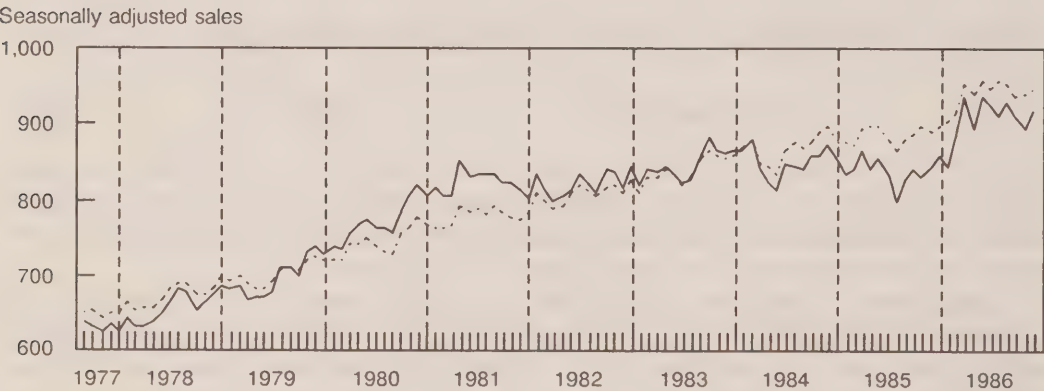


Figure 2.b Drinking Places: Composite (solid) and Signal Extraction (dotted) Estimates

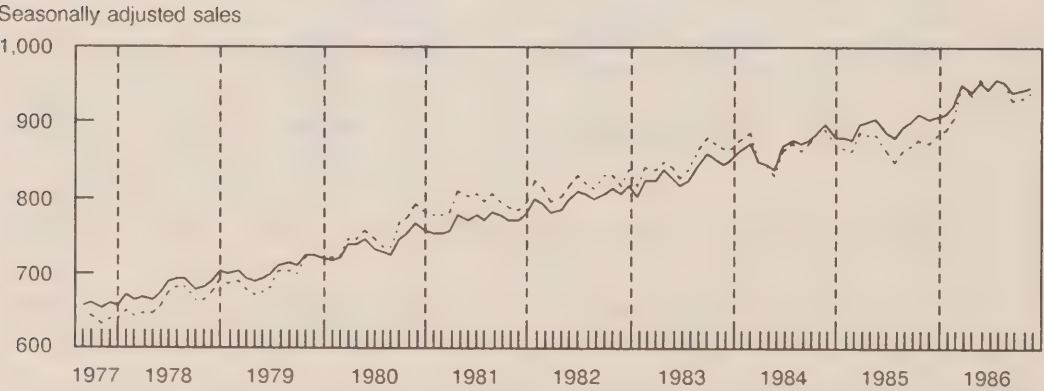


Figure 2.c Drinking Places: Alternative Signal Extraction Estimates

model is known. Examination of standardized residuals produced by the Kalman filter, and of their autocorrelations, suggested no major inadequacies with the fitted models for either series.

The estimated models, (5.7a,b) with (5.9a,b), were used to produce signal extraction estimates of $\log(\theta_t)$, which were then exponentiated to produce estimates of θ_t . The results are shown in Figures 2a,b for the series with the estimated seasonal and trading-day effects removed. Notice that signal extraction makes only slight differences in the estimates for Eating Places, which contained little sampling error (low relative variance), but it makes a considerable difference in the estimates for Drinking Places, which contained much more sampling error (higher relative variance). Signal extraction variances for $\log(\theta_t)$ were also produced; these are relative variances for the estimates of θ_t . Table 2 shows that, depending on the location in the series, signal extraction produces about an 8%–32% improvement in CV over the final composite estimates for Eating Places (though the composite estimate CV is small), and about a 27%–38% improvement in CV for Drinking Places. As noted previously, these results are optimistic, since they assume the true component models are those that were estimated. To partly address concerns about this, we next examine the sensitivity of the results for Drinking Places to variation in the model parameters.

5.3 Sensitivity Analysis for Drinking Places

Here we focus on sensitivity of results to variation in the sampling error model, since this was determined with less information than the signal model. Our approach is to vary parameters of the sampling error model, then reestimate the signal model and redo the signal extraction. While it would be preferable to have more formal statistical measures of the signal extraction error due to model error (which the present state of theory and computer software does not allow), this approach should at least help indicate in what respects the signal extraction results are sensitive to parameter variation and in what respects they are not.

Comparing models (5.8b) and (5.9b) gives some indication of the sensitivity of the signal model to changes in σ_c^2 , the innovation variance of the sampling error model, since (5.8b) corresponds to $\sigma_c^2 = 0$ and (5.9b) to $\sigma_c^2 = 9.3 \times 10^{-5}$. The most noticeable differences are in the estimate of σ_b^2 , which is to be expected, and in the estimate of the seasonal moving average parameter, η_{12} say, which was found to be essentially 1 in obtaining (5.9b). Reestimation of the signal model for other values of σ_c^2 yielded $\hat{\eta}_{12} \geq .99$ as long as $\sigma_c^2 \geq 3.0 \times 10^{-5}$. In light of this, and to simplify presentation of results, we assume $\eta_{12} = 1$ and use a signal model with seasonal indicator variables as in (5.9b).

Figure 2.c. shows (seasonally and trading-day adjusted) signal extraction estimates $\hat{\theta}_t$ corresponding to sampling error models with $(\phi^3, \Phi) = (.564, .614)$ and $(.764, .814)$, and with $\rho = .986$ and $\text{Var}(\log(u_t)) = .00776$ (the relative variance of the Horvitz-Thompson estimates) held fixed. These cover the extremes of $\hat{\theta}_t$ for the sensitivity analysis. The nature of the different estimates $\hat{\theta}_t$ we have generated seems to roughly correspond to the value of $\text{CV}_{56} = [\text{Var}(\log(\hat{\theta}_{56}) - \log(\theta_{56}))]^{1/2}$, the signal extraction coefficient of variation achieved at the middle of the series. (CV_{56} is very close to the lowest value, which is achieved at $t = 53$ – see Table 2.) The lower CV_{56} is, the smoother $\hat{\theta}_t$ is. CV_{56} is 2.78%, 3.28%, and 3.70% for (ϕ^3, Φ) equal to $(.564, .614)$, $(.664, .714)$, and $(.764, .814)$ respectively. Other estimates $\hat{\theta}_t$ we generated lie closest to the signal extraction estimate in Figure 2.b. or 2.c. with the closest CV_{56} .

We now consider the sensitivity of CV_{56} to variations in the sampling error model parameters, beginning with ρ . The only parameter in (5.7b) affected by a change in ρ is η . Table 3 reports the values of η and corresponding values of ρ considered, and the resulting CV_{56} 's. We see CV_{56} is somewhat sensitive to changes in ρ , especially increases: CV_{56} for $\rho = 1$ (3.49) is 6% larger than for $\rho = .985$ (3.28), the value used for (5.7b).

Table 3
Sensitivity of CV_{56}^1 for Drinking Places to Changes in η (Changes in ρ)

η	.00	-.05	-.10	-.15	-.20	-.25
ρ	.9375	.9642	.9792	.9888	.9953	1.000
CV_{56}	3.03	3.12	3.21	3.31	3.40	3.49

¹ CV_{56} is the signal extraction coefficient of variation for $t = 56$ (the middle of the series), expressed as a percentage, *i.e.* the square root of $Var(\log(\hat{\theta}_t) - \log(\theta_t))$ multiplied by 100.

Table 4
Sensitivity of CV_{56} for Drinking Places to Changes in $Var(\log(u_t))^1$ (Changes in σ_c^2)

$Var(\log(u_t))$.00676	.00726	.00776	.00826	.00876
$CV(HT)^2$	8.22	8.52	8.81	9.09	9.36
$\sigma_c^2 \times 10^5$	8.16	8.76	9.30	9.97	10.57
CV_{56}	3.16	3.23	3.28	3.35	3.40

¹ $Var(\log(u_t))$ is the relative variance of the Horvitz-Thompson estimators.
² $CV(HT)$ is the coefficient of variation of the Horvitz-Thompson estimators, expressed as a percentage, *i.e.* the square root of $Var(\log(u_t))$ multiplied by 100.

Table 5
Sensitivity of Results for Drinking Places to Changes in (ϕ^3, Φ)

		(i) Values of $\sigma_c^2 \times 10^5$ for given (ϕ^3, Φ)				
		ϕ^3				
		.564	.614	.664	.714	.764
Φ	.614	16.90	14.70	12.36	9.98	7.64
	.664	15.03	13.00	10.87	8.72	6.62
	.714	13.04	11.23	9.30	7.44	5.60
	.764	10.96	9.40	7.78	6.15	4.58
	.814	8.79	7.51	6.17	4.85	3.58
		(ii) Values of CV_{56} for given (ϕ^3, Φ)				
		ϕ^3				
		.564	.614	.664	.714	.764
Φ	.614	2.78	2.88	2.99	3.12	3.27
	.664	2.95	3.04	3.14	3.26	3.38
	.714	3.10	3.19	3.28	3.39	3.50
	.764	3.24	3.33	3.42	3.51	3.60
	.814	3.36	3.45	3.54	3.62	3.70

We next consider the sensitivity of CV_{56} to changes in $Var(\log(u_t))$. The only sampling error model parameter this affects is σ_c^2 . Table 4 reports the values of $Var(\log(u_t))$, its square root $CV(HT)$, the corresponding σ_c^2 , and the resulting CV_{56} . We see less sensitivity of CV_{56} here than in Table 3.

Finally, we examine the sensitivity of CV_{56} to ϕ^3 and Φ . Holding $\text{Var}(\log(u_t))$ fixed at .00776 and changing (ϕ^3, Φ) also changes σ_c^2 . Table 5 reports the grid of values used for (ϕ^3, Φ) , and resulting values of σ_c^2 and CV_{56} . Notice σ_c^2 varies more here than in Table 4. We see CV_{56} increases substantially as ϕ^3 and Φ are increased.

We conclude from this analysis that moderate changes in the sampling error model parameters have relatively small impacts on $\hat{\theta}_t$. The largest changes we observed in $\hat{\theta}_t$ were around 2 percent. The same moderate changes in the sampling error model parameters have relatively larger impacts on the signal extraction variances, with CV_{56} 's changing by as much as 17 percent. This suggests that for this example the greatest concern in not knowing the sampling error model parameters may be in the effect on signal extraction variances, and the resulting measures of improvement over the composite estimates. However, in all the cases considered in the sensitivity analysis the signal extraction estimates showed a significant improvement in variance.

5.4 Conclusions

The Drinking Places example illustrates the potential gains that may be achieved with the time series approach to survey estimation. Both examples also illustrate the complex and delicate nature of the time series modeling that may be required. We view the results as preliminary for several reasons. First, the optimistic nature of the signal extraction variances that do not reflect parameter estimation error has been mentioned. Second, we have no clear explanation of why the signal extraction estimates lie above or below the composite estimates for long stretches of time. (This is obvious in Figure 2.b., and actually the case in Figure 2.a. as well.) For the Drinking Places example this behavior was evident throughout the sensitivity analysis, and so does not appear to be due to uncertainty in the parameters of the sampling error model. We are in the process of exploring whether this may be due to the forms of the sampling error model or signal model being incorrect. In fact, Bell and Wilcox (1990) report that the correlations of e'_t and e'_{t-1} at lags not multiples of three are not necessarily zero, as was assumed by the model.

ACKNOWLEDGEMENTS

This paper is based in part upon work supported by the National Science Foundation under grant SES 84-01460, "On-Site Research to Improve the Government-Generated Social Science Data Base," and by the U.S. Bureau of the Census and the University of Kansas under Joint Statistical Agreements 87-14 and 88-27. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Census Bureau, or the University of Kansas. We wish to thank Phillip Kott and David Findley for valuable suggestions regarding the consistency results, Abdelwahed Trabelsi for his support as an ASA/NSF/Census Research Associate, and an anonymous referee whose comments have markedly improved the paper. Thanks also to Ruth Detlefsen, Michael Shimberg, and Carol Veum of the Business Division of the Census Bureau for providing data from and information about the Retail Trade Survey. Finally, special thanks to James Bozik, Mark Otto, and Marian Pugh for their extensive work on developing the computer software used in this paper. Any errors in the analysis of the examples are the responsibility of the authors.

REFERENCES

- BELL, W.R. (1984). Signal extraction for nonstationary time series. *Annals of Statistics*, 12, 646-664.
- BELL, W.R. (1987). A Note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with (0,1,1)₁₂ seasonal parts when $\Theta = 1$. *Journal of Business and Economic Statistics*, 5, 383-387.
- BELL, W.R., and HILLMER, S.C. (1983). Modeling time series with calendar variation. *Journal of the American Statistical Association*, 78, 526-534.
- BELL, W.R., and HILLMER, S.C. (1989). Modeling time series subject to sampling error. Research Report 89/01, Statistical Research Division, Bureau of the Census.
- BELL, W.R., and HILLMER, S.C. (1990). A Matrix approach to signal extraction for nonstationary time series models. Submitted for publication.
- BELL, W.R., and WILCOX, D.W. (1990). The effect of sampling error on the time series behavior of consumption data. Paper presented at the CRDE/Journal of Econometrics Conference on Seasonality in Econometric Models, Montreal, Canada, May 1990.
- BINDER, D.A., and DICK, J.P. (1986). Modelling and estimation for repeated surveys. Statistics Canada Technical Report. Social Survey Methods Division.
- BINDER, D.A., and DICK, J.P. (1989). Implications of survey designs for estimating seasonal ARIMA models. Paper presented at the annual meeting of the American Statistical Association, Washington, D.C.
- BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.
- CHUNG, K.L. (1968). *A Course in Probability Theory*. New York: Harcourt, Brace and World, Inc.
- ELTINGE, J.L., and FULLER, W.A. (1989). Time series random component models for sample surveys. Paper presented at the winter meeting of the American Statistical Association, San Diego, CA.
- FULLER, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
- FULLER, W.A., and ISAKI, C.T. (1981). Survey design under superpopulation models. *Current Topics in Survey Sampling*, (Eds. D. Krewski, R. Platek, and J.N.K. Rao). New York: Academic Press, 199-226.
- GARRETT, J.K., DETLEFSEN, R.E., and VEUM, C.S. (1987). Recent sample revisions and related enhancements for business surveys of the U.S. Bureau of the Census. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 141-149.
- GRANGER, C.W.J., and NEWBOLD, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society, Series B*, 38, 189-203.
- HAUSMAN, J.A., and WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- HILLMER, S.C., and TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.
- ISAKI, C.T., WOLTER, K.M., STURDEVANT, T.R., MONSOUR, N.J., and TRAGER, M.L. (1976). Sample redesign of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 90-98.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.
- MCLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.
- MCLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 26, 194.

- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, to appear.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. *Panel Surveys*, (Eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, M.P. Singh). New York: Wiley, 457-479.
- SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F., and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri). New York: Academic Press, 201-216.
- TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.
- TRABELSI, A., and HILLMER, S.C. (1990). Benchmarking time series with reliable benchmarks. *Applied Statistics*, 39, to appear.
- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WOLTER, K.M., ISAKI, C.T., STURDEVANT, T.R., MONSOUR, N.J., and MAYES, F.M. (1976). Sample selection and estimation aspects of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 99-109.
- WOLTER, K.M., and MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (Eds. D. Krewski, R. Platek, and J.N.K. Rao). New York: Academic Press, 199-226.

Robust Small Area Estimation Combining Time Series and Cross-Sectional Data

D. PFEFFERMANN and L. BURCK¹

ABSTRACT

The common approach to small area estimation is to exploit the cross-sectional relationships of the data in an attempt to borrow information from one small area to assist in the estimation in others. However, in the case of repeated surveys, further gains in efficiency can be secured by modelling the time series properties of the data as well. We illustrate the idea by considering regression models with time varying, cross-sectionally correlated coefficients. The use of past relationships to estimate current means raises the question of how to protect against model breakdowns. We propose a modification which guarantees that the model dependent predictors of aggregates of the small area means coincide with the corresponding survey estimators and we explore the statistical properties of the modification. The proposed procedure is applied to data on home sale prices used for the computation of housing price indexes.

KEY WORDS: Kalman filter; Linear constraints; State-space models.

1. INTRODUCTION

Statistical Bureaus are often confronted with the demand to provide reliable estimators for small area means. The problem with the production of such estimators is that the sample sizes within those areas are usually too small to allow the use of direct survey estimators. As a result, new estimators have been proposed in recent years which combine auxiliary information (obtained from a census or administrative records) with the survey data obtained from all the small areas. The common feature of these estimators is that they can be structured in general as a linear combination of two components: a "synthetic estimator" of the form $\bar{X}_i' \hat{\beta}$ where \bar{X}_i represents the average auxiliary information at the small area level and $\hat{\beta}$ is a vector of estimated regression coefficients; and a "correction factor" of the form $(\bar{y}_i - \bar{x}_i' \hat{\beta})$ where \bar{y}_i and \bar{x}_i are the sample means of the target and the auxiliary variables. The correction factors are used to account for the variability of the small area means not explained by the auxiliary variables. The major difference between the various estimators is in the approach followed to determine the weights assigned to the two components in the linear combination, ranging from a "design based approach" (Särndal and Hidiroglou 1989) to "empirical Bayes" (Fay and Herriot 1979) and "mixed linear models" (Battese, Harter and Fuller 1989, Pfeffermann and Barnard 1991).

Very few studies are reported in the literature on the possible use of the time series relationships of the data to further increase the efficiency of the small area estimators. This is despite the fact that many of the small area estimators are derived from repeated surveys such as labour force surveys. The econometric literature contains a vast number of studies on the combined modelling of time series and cross-sectional data, see *e.g.* Rosenberg (1973b), Johnson (1977, 1980), Maddala (1977, Chapter 7), Dielman (1983) and Pfeffermann and Smith (1985) for reviews. However, none of these studies is directed to the problem of estimating (predicting) small area means from survey data. Fitting time series models to survey data has been considered

¹ D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905. L. Burck, Unit for Statistical Analysis, Central Bureau of Statistics, Jerusalem 91130.

in the context of estimating aggregate population means, see the review papers of Smith (1979) and Binder and Hidirolou (1988) and the more recent articles by Binder and Dick (1989), Tiller (1989) and Pfeffermann (1991). But again, these methods are not in routine use mainly because the classical survey estimators of the aggregate means are often almost as efficient when the models hold and more robust when the models fail to hold.

The situation is clearly different when dealing with a small area estimation problem; it seems to us that for this kind of problem, the use of time series models can be of great advantage. Although the exact nature of the model to be used in a particular application is obviously 'data dependent', the class of models we consider in the next section is broad enough to apply to many, if not most of the small area estimation problems arising in practice. These models have the further advantage that their estimation is relatively simple. Estimation issues are discussed in Section 3.

The use of a model always raises the question of how to protect against possible model failures and this question becomes even more sensitive when considering the use of a model for the production of official statistics. In Section 4 we consider this issue and propose a modification to the model dependent predictors which guarantees that for aggregates of the small area means for which the direct survey estimators can be trusted, the modified model predictors coincide with the survey estimators. The statistical properties of the modified predictors are explored. We conclude the article in Section 5 with empirical results which illustrate the performance of the model with and without the proposed modification. The data used for the illustrations are the sale prices of homes in the city of Jerusalem during the months of September 1985 through November 1989. These data are used routinely by the Central Bureau of Statistics in Israel for the computation of housing price indexes.

2. REGRESSION WITH CROSS-SECTIONALLY AND TIME VARYING COEFFICIENTS

2.1 A General Class of Models

In what follows we denote by \underline{Y}_{tk} the $n_{tk} \times 1$ vector of observations on a target variable Y , pertaining to an area k at time t , $k = 1, \dots, K$, $t = 1, 2, \dots$. We assume for convenience that $n_{tk} \geq 1$ but as becomes evident later on, the model permits that some of the areas not be observed at certain times. Let X_{tk} define the corresponding $n_{tk} \times (p + 1)$ design matrix of the auxiliary variables with a vector of ones as its first column. In many applications, the same row vector \underline{x}'_{ik} of auxiliary values applies to all the Y values of a given time so that $X_{tk} = \underline{1}_{n_{tk}} \underline{x}'_{ik}$ where $\underline{1}_{n_{tk}}$ is a column vector of ones of length n_{tk} . This is the case when the only available data are the small area survey estimators. Confidentiality as well as processing costs often preclude the use of micro data on individual survey respondents. The theory described in this article is not restricted to the availability of the micro data (see the example in Section 2.2) but data availability has an obvious effect on model specifications and precision of estimation.

The regression model holding in area k at time t is defined as

$$\underline{Y}_{tk} = X_{tk} \underline{\beta}_{tk} + \underline{\epsilon}_{tk}; E(\underline{\epsilon}_{tk}) = 0, E(\underline{\epsilon}_{tk} \underline{\epsilon}'_{tk}) = \sigma_k^2 I_{n_{tk}} \quad (2.1)$$

where $\underline{\beta}_{tk} = (\beta_{tk0}, \beta_{tk1}, \dots, \beta_{tkp})$.

We define the (superpopulation) mean of the target variable values in area k at time t to be

$$\Theta_{tk} = E(M_{tk} | \underline{\beta}_{tk}) = \bar{X}_{tk} \underline{\beta}_{tk} \quad (2.2)$$

where

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} \quad \text{and} \quad \bar{X}_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} x'_{tki}$$

with $i = 1, \dots, N_{tk}$ indexing the population units. Obviously, when $x'_{tki} \equiv x'_{tk}$, then $\bar{X}_{tk} = x'_{tk}$.

Let $\hat{\beta}_{tk}$ define an estimator for β_{tk} . Then $\hat{\Theta}_{tk} = \bar{X}_{tk} \hat{\beta}_{tk}$ and

$$\hat{M}_{tk} = \frac{1}{N_{tk}} \left[\sum_{i=1}^{n_{tk}} Y_{tki} + \sum_{i=n_{tk}+1}^{N_{tk}} x'_{tki} \hat{\beta}_{tk} \right] = \hat{\Theta}_{tk} + \frac{1}{N_{tk}} \left(\sum_{i=1}^{n_{tk}} (Y_{tki} - x'_{tki} \hat{\beta}_{tk}) \right)$$

implying that in the usual case of small sampling rates within the areas, $\hat{\Theta}_{tk}$ can also be considered as an estimator of the finite population mean M_{tk} . For this reason we no longer distinguish between the finite and superpopulation means.

The notable feature of (2.1) is that the coefficients β_{tk} are allowed to vary both cross-sectionally and over time. The following equations specify the variation of the coefficients over time:

$$\begin{bmatrix} \beta_{tkj} \\ \beta_{kj} \end{bmatrix} = T_j \begin{bmatrix} \beta_{t-1,kj} \\ \beta_{kj} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{tkj}, \quad j = 0, \dots, p \quad (2.3)$$

where we use the notation β_{kj} , $j = 0, 1, \dots, p$, to define fixed coefficients which we interpret below, and T_j to define fixed (2×2) matrices and where the residuals $\{\eta_{tkj}\}$ satisfy

$$E(\eta_{tkj}) = 0, \quad E(\eta_{tkj} \eta_{tkl}) = \delta_{jl}, \quad E(\eta_{tkj} \eta_{t-d,k\ell}) = 0 \quad \text{for } d > 0. \quad (2.4)$$

The implication of (2.4) is that residuals of different coefficients pertaining to the same time t are allowed to be correlated but the serial and cross serial correlations are assumed to be zero.

Next, we illustrate the use of (2.3) by considering some simple cases:

- (a) $T_j = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ implies that $\beta_{tkj} = \beta_{kj} + \eta_{tkj}$ so that β_{kj} represents, in this case, a common mean. This is the well known Random Coefficient Regression Model (Swamy 1971) which is often used in econometric applications. Obviously, by postulating, $\text{var}(\eta_{tkj}) = 0$, the model reduces to the case of a fixed regression coefficient over time.
- (b) $T_j = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$ which is the familiar random walk model, see *e.g.* Cooley and Prescott (1976) and LaMotte and McWhorter (1977) for application of this model in econometric studies. In this case the coefficient β_{kj} is redundant and should be omitted so that $T_j \equiv 1$.
- (c) $T_j = \begin{bmatrix} \rho & 1-\rho \\ 0 & 1 \end{bmatrix}$ implies the first order autoregressive relationship $(\beta_{tkj} - \beta_{kj}) = \rho(\beta_{t-1,kj} - \beta_{kj}) + \eta_{tkj}$ considered by Rosenberg (1973a).
- (d) $T_j = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ implies that $\beta_{tkj} = \beta_{t-1,kj} + \beta_{kj} + \eta_{tkj}$ which defines a local approximation to a linear trend (Kitagawa and Gersch 1984). The coefficient β_{kj} represents, in this case, a fixed slope.

It should be emphasized that different matrices T_j can be used for different coefficients β_{tkj} . In fact, by defining $\alpha'_{tk} = (\beta_{tk0}, \beta_{tk0}, \beta_{tk1}, \beta_{tk1}, \dots, \beta_{tkp}, \beta_{tkp})$; $\tilde{T} = \text{diag}[T_0, T_1, \dots, T_p]$, a block diagonal matrix with T_j as the j -th block; $\tilde{G} = I_{p+1} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ where I_{p+1} is the identity matrix of order $p + 1$ and \otimes defines the Kronecker product and $\eta'_{tk} = (\eta_{tk0}, \eta_{tk1}, \dots, \eta_{tkp})$, the combined model holding for the coefficients β_{tk} can be written as

$$\alpha_{tk} = \tilde{T}\alpha_{t-1,k} + \tilde{G}\eta_{tk}; \quad E(\eta_{tk}) = 0, E(\eta_{tk}\eta'_{t-d,k}) = A_d\Delta \quad (2.5)$$

where $A_d = 1$ for $d = 0$ and $A_d = 0$ otherwise, and $\Delta = [\delta_{ij}]$ is defined by the variances and covariances δ_{ij} (equation 2.4).

The model defined by (2.5) specifies the variation of the regression coefficients of a specific area over time. The common approach to account for cross-sectional relationships between small area means is to allow for random small area effects which are time invariant $\{u_k\}$. The general model defined by (2.1) and (2.3) includes this case by writing $Y_{tk} = 1_{n_{tk}}u_{tk} + X_{tk}\beta_{tk} + \epsilon_{tk} = X_{tk}^*\beta_{tk}^* + \epsilon_{tk}$, say, and specifying $u_{tk} = u_{t-1,k} + \eta_{tk}$ with $u_{0k} = 0$, $\text{var}(\eta_{1k}) = \sigma_\eta^2$ and $\text{var}(\eta_{tk}) = 0$ for $t > 1$ (compare with case (b) above). By assuming in addition the autoregressive relationship defined by case (c) for the intercept variable and fixing the other regression coefficients (case (a) with zero residual variances), the resulting model is similar to the model considered by Choudhry and Rao (1989) except that in their general formulation of the model the observation residuals of equation (2.1) are allowed to be serially correlated. Notice that equation (2.1) now contains two random "intercept terms" but the model is nonetheless identifiable. Choudhry and Rao assume that the only available data are the survey estimators so that the estimation of the serial correlations needs to be carried out externally, using the micro observations. Alternatively, a model accounting for the serial correlations can be postulated. Choudhry and Rao assume an AR(1) model in their study.

A more general way to account for the cross-sectional relationships between the small area means is to allow for non zero correlations between the residual terms η_{tkj} and η_{tmj} of the models specifying the time series variation of the regression coefficients β_{tkj} and β_{tmj} operating in areas k and m (equation 2.4). Often it is reasonable to assume that the correlations decay as the distance between the areas increases. This can be formulated as, $E(\eta_{tkj}, \eta_{tmj}) = \delta_{jj}\rho_j f_j(k, m)$, $k \neq m$, where $f_j(k, m)$ is a monotonic decreasing function of the distances $D(k, m)$. The case of geometrically decaying correlations is obtained by defining $f_j(k, m) = \rho_j^{|k-m|-1}$. The case of fixed correlations is obtained by specifying $f_j(k, m) \equiv 1$ and in what follows we consider this case only. Allowing for fixed cross-sectional correlations for all the regression coefficients can be formulated as

$$E(\eta_{tk}\eta'_{tm}) = D(\Delta)\theta, \quad k \neq m \quad (2.6)$$

where $D(\Delta)$ is the diagonal matrix with the variances δ_{jj} on the main diagonal and θ is another diagonal matrix composed of the correlations ρ_j .

Before concluding this section we present the model defined by (2.1), (2.5) and (2.6) in a state-space form. Presenting the model in this form has important computational advantages.

Let $Y'_t = (Y'_{t1}, \dots, Y'_{tK})$ represent the vector of observations of length $n_t = \sum_k n_{tk}$ for all the areas at time t and let $\epsilon'_t = (\epsilon'_{t1}, \dots, \epsilon'_{tK})$ represent the corresponding regression residuals. Define $Z_{tk} = [1_{n_{tk}}, Q_{ntk}, X_{tk1}, Q_{ntk}, \dots, X_{tkp}, Q_{ntk}]$ where Q_{ntk} is a vector of zeroes of length n_{tk} and X_{tkj} is the vector of values for the j -th auxiliary variable, $j = 1, \dots, p$. Let Z_t be the block diagonal matrix composed of the matrices Z_{tk} . The matrix Z_t is of order $n_t \times [K \times 2 \times (p + 1)]$. Define also $\alpha'_t = (\alpha'_{t1}, \dots, \alpha'_{tK})$, $\eta'_t = (\eta'_{t1}, \dots, \eta'_{tK})$, $\Sigma_t = \text{Diag}[\sigma_1^2 1'_{n_{t1}}, \dots, \sigma_K^2 1'_{n_{tK}}]$, $T = I_K \otimes \tilde{T}$, and $G = I_K \otimes \tilde{G}$.

Using this notation, the model defined by (2.1), (2.5) and (2.6) can be written compactly as

$$\underline{Y}_t = \underline{Z}_t \underline{\alpha}_t + \underline{\epsilon}_t; E(\underline{\epsilon}_t) = \underline{0}, E(\underline{\epsilon}_t \underline{\epsilon}_t') = \underline{\Sigma}_t \quad (2.7)$$

$$\underline{\alpha}_t = \underline{T} \underline{\alpha}_{t-1} + \underline{G} \underline{\eta}_t; E(\underline{\eta}_t) = \underline{0}, E(\underline{\eta}_t \underline{\eta}_t') = \underline{\Lambda}, \quad (2.8)$$

where $\underline{\Lambda} = [\Lambda_{k\ell}]$, $k, \ell = 1, \dots, K$ with $\Lambda_{k\ell} = \Delta$ when $k = \ell$ and $\Lambda_{k\ell} = D(\Delta)\emptyset$ when $k \neq \ell$. The matrices $\Lambda_{k\ell}$ are $(p+1) \times (p+1)$.

The model defined by (2.7) and (2.8) conforms to the classical state-space formulation, see, *e.g.* Anderson and Moore (1979) and Harvey (1984). By this formulation, (2.7) is the observation equation and (2.8) is the state equation with $\underline{\alpha}_t$ defining the state vector. The apparent advantage of restructuring the model in a state space form is that the vectors $\underline{\alpha}_t$, and hence the population means Θ_{tk} , as well as the estimation error variances can be estimated conveniently by means of the Kalman filter. We discuss the use of the filter in sections 3 and 4.

2.2 Explicit Estimators of the Small Area Means

In order to illustrate how past and neighbouring data are used under the model to "strengthen" the small area estimators we consider the case where the same vector \underline{x}_{tk} of auxiliary values applies to all the units of a given area at a given time. In this case the observation equation can be formulated in terms of the sample means, *i.e.*

$$\bar{Y}_{tk} = \underline{x}'_{tk} \underline{\beta}_{tk} + \bar{\epsilon}_{tk}; E(\bar{\epsilon}_{tk}) = 0, E(\bar{\epsilon}_{tk}^2) = \sigma_k^2/n_{tk}, k = 1, \dots, K. \quad (2.9)$$

Suppose that the regression coefficients follow a random walk (case (b) of equation 2.3) so that for area k

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}; E(\eta_{tkj}) = 0, E(\eta_{tkj} \eta_{tk\ell}) = \delta_{j\ell}, j, \ell = 1, \dots, p \quad (2.10)$$

and for areas $k \neq m$,

$$E(\eta_{tkj} \eta_{tmj}) = \delta_{jj} \rho_j; E(\eta_{tkj} \eta_{tm\ell}) = 0, j \neq \ell. \quad (2.11)$$

The random walk model implies that the coefficients drift slowly away from their initial value with no inherent tendency to return to a mean value. Obviously, for residuals η_{tkj} such that $E(\eta_{tkj}^2) = 0$ the corresponding regression coefficients are fixed over time. Notice also that since $\underline{\beta}_{tk} = \underline{\beta}_{t-1,k} + \underline{\eta}_{tk}$, the predictor of $\underline{\beta}_{tk}$ at time $(t-1)$ is the same as the predictor $\hat{\underline{\beta}}_{t-1,k}$ of $\underline{\beta}_{t-1,k}$.

Using the Kalman filter equations presented in section 3, it is shown in the Appendix that the estimator $\hat{\Theta}_{tk}$ of the small area mean Θ_{tk} (equation 2.2) can be structured in this case in the following form

$$\hat{\Theta}_{tk} = \underline{x}'_{tk} \hat{\underline{\beta}}_{t-1,k} + \left(1 - \frac{\sigma_k^2}{n_{tk} v_k^2}\right) (\bar{Y}_{tk} - \underline{x}'_{tk} \hat{\underline{\beta}}_{t-1,k}) + \frac{\sigma_k^2}{n_{tk} v_k^2} \sum_{\substack{m=1 \\ m \neq k}}^K \gamma_{km} (\bar{Y}_{tm} - \underline{x}'_{tm} \hat{\underline{\beta}}_{t-1,m}) \quad (2.12)$$

where the coefficients $\{\gamma_{km}\}$ are the partial regression coefficients in the regression of $e_{tk} = (\bar{Y}_{tk} - x'_{tk} \hat{\beta}_{t-1,k})$ against the prediction errors $\{e_{tm} = (\bar{Y}_{tm} - x'_{tm} \hat{\beta}_{t-1,m})\}$ obtained in the other areas and v_k^2 is the residual (unexplained) variance in the regression.

The estimator $\hat{\Theta}_{tk}$ is composed of three components: the "synthetic" estimator, $x'_{tk} \hat{\beta}_{t-1,k}$, where $\hat{\beta}_{t-1,k}$ is the optimal predictor of β_{tk} based on all the observations up to and including time $t - 1$, the "correction factor" $(\bar{Y}_{tk} - x'_{tk} \hat{\beta}_{t-1,k})$ based on the prediction error in area k , and an "adjustment factor" based on the prediction errors observed for the other areas. The first two components correspond to the components of the classical small area estimators discussed in the introduction. Notice that the smaller the sample size n_{tk} , the smaller is the weight assigned to the current sample mean \bar{Y}_{tk} in the estimation of Θ_{tk} and the larger is the weight assigned to the time series predictor $x'_{tk} \hat{\beta}_{t-1,k}$. The third component in the right hand side of (2.12) represents the information borrowed from neighbouring areas. The weight assigned to this component depends on the magnitude of the correlations ρ_j between the corresponding error terms $\{\eta_{tkj}\}$ in the models holding for the regression coefficients (equation 2.11). Obviously, when the regressions in the various areas are independent so that $\rho_j = 0$ for all j and hence $\gamma_{km} = 0$ for all m , the third component vanishes and the predictor $\hat{\Theta}_{tk}$ reduces to a weighted average of the current mean \bar{Y}_{tk} and the time series predictor $x'_{tk} \hat{\beta}_{t-1,k}$.

3. MODEL ESTIMATION AND INITIALIZATION USING THE KALMAN FILTER

3.1 Estimation of the Regression Coefficients by Means of the Kalman Filter

In this section we present the Kalman filter equations for the updating and smoothing of the state vectors α_t defined by the equations (2.7) and (2.8) (the area regression coefficients in our case). We assume that the V-C matrices Σ_t and Λ are known. Estimation of these matrices is considered in section 3.2. The theory of the Kalman filter is developed in numerous publications (see e.g. Anderson and Moore 1979 and Meinhold and Singpurwalla 1983) and so we restrict the discussion to aspects most germane to the small area estimation problem.

Let $\hat{\alpha}_{t-1}$ be the best linear unbiased predictor (blup) of α_{t-1} based on all the data observed up to time $(t - 1)$. Since $\hat{\alpha}_{t-1}$ is blup for α_{t-1} , $\hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1}$ is the blup of α_t at time $(t - 1)$. Furthermore, if $P_{t-1} = E(\hat{\alpha}_{t-1} - \alpha_{t-1})(\hat{\alpha}_{t-1} - \alpha_{t-1})'$ is the V-C matrix of the prediction errors at time $(t - 1)$, $P_{t|t-1} = TP_{t-1}T' + GAG'$ is the V-C matrix of the prediction errors $(\hat{\alpha}_{t|t-1} - \alpha_t)$. (Follows straightforwardly from 2.8).

When a new vector of observations $[Y_t, Z_t]$ becomes available, the predictor of α_t and the V-C matrix P_{t-1} are updated according to the formulae

$$\begin{aligned}\hat{\alpha}_t &= \hat{\alpha}_{t|t-1} + P_{t|t-1}Z_t'F_t^{-1}(Y_t - \hat{Y}_{t|t-1}) \\ P_t &= (I - P_{t|t-1}Z_t'F_t^{-1}Z_t)P_{t|t-1}\end{aligned}\tag{3.1}$$

where $\hat{Y}_{t|t-1} = Z_t\hat{\alpha}_{t|t-1}$ is the blup of Y_t at time $(t - 1)$ so that $e_t = (Y_t - \hat{Y}_{t|t-1})$ is the vector of innovations with V-C matrix $F_t = (Z_tP_{t|t-1}Z_t' + \Sigma_t)$.

The new data observed at time t can be used also for the updating (smoothing) of past estimators of the state vectors and hence for the updating of past estimators of the small area means. Denoting by t^* the most recent month with observations, the smoothing is carried out using the equations

$$\hat{\alpha}_{t|t^*} = \hat{\alpha}_t + P_t T' P_{t+1|t}^{-1} (\hat{\alpha}_{t+1|t^*} - T \hat{\alpha}_t) \quad (3.2)$$

$$P_{t|t^*} = P_t + P_t T' P_{t+1|t}^{-1} (P_{t+1|t^*} - P_{t+1|t}) P_{t+1|t}^{-1} T P_t; \quad t = 2, 3, \dots, t^*$$

where $P_{t|t^*}$ is the V-C matrix of the prediction errors $(\hat{\alpha}_{t|t^*} - \alpha_t)$. Notice that $\hat{\alpha}_{t^*|t^*} = \hat{\alpha}_{t^*}$ and $P_{t^*|t^*} = P_{t^*}$ define the starting values for the smoothing equations.

Estimators of the small area means or aggregates of the means are obtained from the filtered (or smoothed) estimators of α_t in a straightforward manner using the relationship $\hat{\Theta}_{tk} = \bar{X}_{tk} \hat{\beta}_{tk} = \bar{Z}_{tk}' \hat{\alpha}_{tk} = \bar{Z}_{tk}' A_{tk} \hat{\alpha}_t$ where $\bar{Z}_{tk}' = (1, 0, \bar{X}_{tk1}, 0, \dots, \bar{X}_{tkp}, 0)$ and A_{tk} is the appropriate indicator matrix. Hence, if $\Theta_t^w = \sum_{k=1}^K w_k \Theta_{tk}$, then $\hat{\Theta}_t^w = \sum_{k=1}^K w_k \bar{Z}_{tk}' A_{tk} \hat{\alpha}_t = \underline{q}_{tw}' \hat{\alpha}_t$, say. For given V-C matrices Σ_t and Λ , the MSE's of the estimation errors are obtained as

$$E(\hat{\Theta}_{tk} - \Theta_{tk})^2 = \bar{Z}_{tk}' A_{tk} P_t A_{tk}' \bar{Z}_{tk} \quad \text{and} \quad E(\hat{\Theta}_t^w - \Theta_t^w) = \underline{q}_{tw}' P_t \underline{q}_{tw}. \quad (3.3)$$

Notice that the MSE's in (3.3) are with respect to the joint distribution of the observations $\{Y_{tk}\}$ and the vectors of coefficients $\{\beta_{tk}\}$ so that they represent average MSE's over the possible realizations of the area means.

3.2 Estimation of the V-C Matrices and Initialization of the Filter

The actual application of the Kalman filter requires the estimation of the unknown elements of the matrices Σ_t and Λ and the initialization of the filter, that is, the estimation of the vector α_o and the corresponding V-C matrix P_o of the estimation errors. In this section we describe simple estimation procedures which can be used for these purposes.

Assuming a normal distribution for the residual terms ϵ_t and η_t of equations (2.7) and (2.8), the log likelihood function of the vectors Y_{m+1}, \dots, Y_{t^*} , conditional on the first m vectors Y_1, \dots, Y_m , can be formulated as

$$L(\lambda) = \text{constant} - \frac{1}{2} \sum_{t=m+1}^{t^*} (\log |F_t| + \underline{e}_t' F_t^{-1} \underline{e}_t) \quad (3.4)$$

where λ contains the unknown model variances and covariances written in a vector form. The scalar m defines the number of time periods needed to construct initial values for the Kalman filter. (For the random walk model considered in section 2.2, $m = 1$, provided that sufficient data are available in every area to allow the computation of the OLS estimators of the vectors of coefficients). The expression in (3.4) follows from the prediction error decomposition, see Schweppe (1965) and Harvey (1981) for details. For given matrices Σ_t and Λ , the innovations \underline{e}_t and the V-C matrices F_t can be obtained by application of the Kalman filter equations (3.1).

The computation of the likelihood function requires the initialization of the Kalman filter which can be carried out most conveniently by application of the approach proposed by Harvey and Phillips (1979). By this approach, the nonstationary components of the state vector are initialized with very large error variances which corresponds to postulating a noninformative prior distribution so that the corresponding state estimates can conveniently be taken as zeroes. (For the random walk model, initializing with a noninformative prior yields the OLS estimators after one time period, see Meinhold and Singpurwalla 1983, for a Bayesian formulation of the Kalman filter). The stationary components of the state vector are initialized by the corresponding unconditional means and variances which may be part of the unknown parameters defining the arguments of the likelihood function.

Maximization of the likelihood function (3.4) can be implemented using the method of scoring with a variable step length. In particular, let $\lambda_{(0)}$ define initial estimates of the unknown elements in λ . Then the method of scoring consists of solving iteratively the set of equations

$$\lambda_{(i)} = \lambda_{(i-1)} + r_i \{I[\lambda_{(i-1)}]\}^{-1} g[\lambda_{(i-1)}] \quad (3.5)$$

where $\lambda_{(i-1)}$ is the estimator of λ as obtained in the $(i - 1)$ -th iteration, $I[\lambda_{(i-1)}]$ is the information matrix evaluated at $\lambda_{(i-1)}$ and $g[\lambda_{(i-1)}]$ is the gradient of the log likelihood evaluated at $\lambda_{(i-1)}$. The coefficient r_i is a variable step length introduced to guarantee that $L[\lambda_{(i)}] \geq L[\lambda_{(i-1)}]$ in every iteration. The value of r_i can be determined by a grid search procedure in the region $[0, 1]$. The formulae for the k -th element of the gradient vector and the $k\ell$ -th element of the information matrix are given in Watson and Engle (1983).

Having estimated the model variances and covariances, these estimates can be substituted for the true parameters in the Kalman filter equations (3.1) – (3.2) to yield the estimators of the regression coefficients and the V-C matrices and hence the small area estimators and their variances (see equation 3.3). Notice however that the estimated V-C matrices ignore the variability induced by the need to estimate the unknown elements contained in λ . Ansley and Kohn (1986) propose correction factors of order $1/t^*$ to account for this extra variation in state space modelling using first order Taylor approximations. Hamilton (1986) proposes a Monte Carlo procedure which consists of sampling from a multivariate normal distribution with mean given by the maximum likelihood estimator of the vector λ and V-C matrix defined by the inverse of the information matrix, and estimating the state vectors for each random realization of the parameter values. This procedure is more flexible in terms of the assumptions involved and provides further insight into the sensitivity of the Kalman filter estimators to errors in the variance and covariance estimators. However, it is computationally more intensive.

4. MODIFICATIONS TO PROTECT AGAINST MODEL BREAKDOWNS

4.1 Description of the Problem and Proposed Modifications

The use of a model for small area estimation seems inevitable in view of the small sample sizes within the areas. However it raises the question of how to protect against model breakdowns. Testing the model every time that new data becomes available is often not practical, requiring instead the development of a “built-in mechanism” to ensure the robustness of the estimators when the model fails to hold.

One possibility is to modify the regression estimators derived in the various time periods so that they satisfy certain linear constraints obtained by equating aggregate means of the raw data with their expected fitted values under the model. More precisely, we propose to augment the model equation (2.1) by linear constraints of the form

$$\sum_k W_{tk}^{(\ell)} \sum_i Y_{tki} = \sum_k W_{tk}^{(i)} \sum_i x'_{tki} \beta_{tk} \quad \ell = 1, 2, \dots, L(t), \quad t = 1, \dots, t^* \quad (4.1)$$

where the coefficients $W_{tk}^{(\ell)}$ are fixed, standardized weights such that $\sum_k n_{tk} W_{tk}^{(\ell)} = 1$. An example for such a constraint would be the equation

$$\sum_{k=1}^K N_{tk} \hat{M}_{tk} \bigg/ \sum_{k=1}^K N_{tk} = \sum_{k=1}^K N_{tk} (\bar{x}_{tk}' \beta_{tk}) \bigg/ \sum_{k=1}^K N_{tk} \quad (4.2)$$

where \hat{M}_{tk} is the direct, survey estimator in area k . For $\bar{x}_{tk} \approx \bar{X}_{tk}$, the equation (4.2) guarantees that the model dependent predictor of the aggregate population mean coincides with the corresponding survey estimator. Such a constraint can be justified by arguing that the survey estimators, although not reliable enough for estimating the small area means due to the small sample sizes, can be trusted when being combined for estimating the aggregate mean. Notice that "adding up" constraints are ordinarily imposed on statistical agencies anyway. Battese, Harter and Fuller (1988) and Pfeiffermann and Barnard (1991) use a similar constraint for analysing cross-sectional surveys. Often, the small areas can be grouped into broader groups, with sufficient data in each of the groups to justify the use of the survey estimators for estimating the corresponding group means. In this case, one can impose several constraints of the form (4.2) where the summation is now over the areas belonging to the same group. Notice in this respect that in view of the correlations between the regression coefficients operating in the various areas, a constraint applied to a sub-set of the areas will modify the regression estimates in all the areas. We illustrate this property in the empirical study.

It is important to emphasize that the set of constraints in (4.1) does not represent external information about possible values of the regression coefficients. Rather, it serves as a "control system" to guarantee that the model estimators adjust themselves more rapidly to possible changes in the behavior of the regression coefficients. As a result, the variances of the modified regression estimators are slightly larger than the variances of the optimal estimators under the model. Obviously, when no such changes occur and the variances of the aggregate means are sufficiently small, one would expect the constraints to be satisfied approximately even without imposing them explicitly. As mentioned above, it is possible to incorporate several separate constraints in each time period but it is imperative that the variances of the corresponding aggregate means will be small enough to ensure that the modifications are indeed needed and do not interfere with the random fluctuation of the raw data.

4.2 Inference Incorporating the Linear Constraints

In Section 4.1 we proposed to amend the model equations (2.1) by imposing the set of constraints (4.1) thereby ensuring the robustness of the regression estimators against sudden drifts in the values of the coefficients.

Computationally, this can be implemented most conveniently by augmenting the vectors Y_t of equation (2.7) by the scalars $\sum_k W_{tk}^{(t)} \sum_i Y_{tki}$, augmenting the matrices Z_t by the corresponding row vectors $(W_{t1}^{(t)} 1'_{n1} Z_{t1}, \dots, W_{tK}^{(t)} 1'_{nK} Z_{tK})$ and setting the respective variances of the residual terms to zero. The augmented set of equations, together with (2.8), form a pseudo state-space model which could be estimated using the Kalman filter equations (3.1). Notice that the pseudo V-C matrix $\Sigma_t^{(P)}$ of the augmented residual vector is no longer positive definite (the last $L(t)$ rows and columns of $\Sigma_t^{(P)}$ consist of zeroes) but this does not cause computational difficulties.

The drawback of applying the Kalman filter to the pseudo model is that the V-C matrices of the regression estimators fail to account for the actual variability of the aggregate means appearing in the left hand side of (4.1). In order to deal with this problem, we propose to amend the formula for the updating of the V-C matrix P_t (equation 3.1) so that the variances and covariances of the aggregate means will be taken into account.

Let $Y_t^{(A)}$ and $Z_t^{(A)}$ represent the augmented Y vector and Z matrix at time t and denote by $\Sigma_t^{(A)}$ the actual V-C matrix of the residual terms $[Y_t^{(A)} - Z_t^{(A)}\alpha_t]$. The matrix $\Sigma_t^{(A)}$ is of order $[n_t + L(t)]$ with Σ_t in the first n_t rows and columns and the variances and covariances of the means $\sum_k W_{tk}^{(\theta)} \sum_i Y_{tki}$ among themselves and with the vector Y_t in the remaining rows and columns. Denoting by $\hat{\alpha}_{t-1}^{(A)}$ the robust predictor of α_{t-1} as obtained at time $(t - 1)$ using the pseudo model and by $P_{t-1}^{(A)}$ the actual V-C matrix of the errors $(\hat{\alpha}_{t-1}^{(A)} - \alpha_{t-1})$, the modified state estimator at time t is obtained as

$$\hat{\alpha}_t^{(A)} = T\hat{\alpha}_{t-1}^{(A)} + P_{t|t-1}^{(A)}Z_t^{(A)'}(F_t^{(P)})^{-1}[Y_t^{(A)} - Z_t^{(A)}T\hat{\alpha}_{t-1}^{(A)}]$$

(4.3)

where $P_{t|t-1}^{(A)} = (TP_{t-1}^{(A)}T' + GAG')$ and $F_t^{(P)} = Z_t^{(A)}P_{t|t-1}^{(A)}Z_t^{(A)'} + \Sigma_t^{(P)}$ (Compare with 3.1). It is shown in the Appendix that the actual V-C matrix $P_t^{(A)}$ of the errors $(\hat{\alpha}_t^{(A)} - \alpha_t)$ satisfies the recursive equation

$$P_t^{(A)} = [I - K_t^{(P)}Z_t^{(A)}]P_{t|t-1}^{(A)} + K_t^{(P)}[\Sigma_t^{(A)} - \Sigma_t^{(P)}]K_t^{(P)},$$

(4.4)

where $K_t^{(P)} = P_{t|t-1}^{(A)}Z_t^{(A)'}(F_t^{(P)})^{-1}$ is the pseudo Kalman gain. The first expression on the right hand side of (4.4) corresponds to the usual updating formula of the Kalman filter (compare with 3.1)). The second expression is a correction factor which accounts for the actual variances and covariances of the means $\sum_k W_{tk}^{(\theta)} \sum_i Y_{tki}$, not taken into account in the first expression.

The amended Kalman filter defined by the equations (4.3) and (4.4) produces robust predictors $\hat{\alpha}_t^{(A)}$ instead of the optimal, model dependent predictors, $\hat{\alpha}_t$ but otherwise uses the correct V-C matrices under the model. Thus, this filter can be used for the routine estimation of the vectors of coefficients and hence for the estimation of the small area means, and when the model holds it will give similar results to those obtained under the optimal filter. In periods where the model fails to hold, the updating formula (4.4) could be incorrect (depending on the particular model failures) but the predictors $\hat{\alpha}_t^{(A)}$ will nonetheless satisfy the linear constraints (4.1). The smoothing equations (3.2) can likewise be modified to satisfy the linear constraints.

5. EMPIRICAL RESULTS

5.1 Description of the Data and Model Fitted

In order to illustrate the important features of the class of models defined in Section 2, we fitted such a model to home sale prices in Jerusalem. The sale prices are recorded on a monthly basis and are routinely used by the Central Bureau of Statistics in Israel for the computation of monthly housing price indexes (HPI) adjusted for changes in quality. The HPI is computed separately for each city or group of cities and for each house size defined by the number of rooms, ranging from 1 to 5. The number of transactions carried out each month is very small in many of these cells and for 1 room apartments it occasionally happens that there are no transactions. The mean and standard deviation (S.D.) of the monthly number of transactions carried out during the period July 1987 – November 1989 are listed below.

Size	1	2	3	4	5
Mean	2.7	29.0	101.9	39.7	5.6
S.D.	2.6	12.9	50.4	18.8	3.5

The need to adjust for changes in quality results from the fact that the transactions performed are not under control, giving rise to large differences in quality from one month to the other particularly in the small cells. The following quality measure variables (QMV) are recorded for every transaction: $\bar{X}^{(1)}$ – the apartment floor area, $\bar{X}^{(2)}$ – the age of the apartment, $X^{(3)}$, $X^{(4)}$ – dummy variables defining districts within the city.

The problems involved in the computation of the HPI and the method used in Israel are discussed at length in a recent article by Pfeffermann, Burck and Ben-Tuvia (1989). The following model was proposed by the authors as an alternative to the model in current use. The triple index “ tki ” defines the i -th transaction of size k in month t with Y_{tki} standing for the log of the sale price and $X_{tki}^{(j)} = \log(\bar{X}_{tki}^{(j)})$, $j = 1, 2$.

$$Y_{tki} = \beta_{tk0} + \beta_{tk1}X_{tki}^{(1)} + \beta_{tk2}X_{tki}^{(2)} + \beta_{tk3}X_{tki}^{(3)} + \beta_{tk4}X_{tki}^{(4)} + \epsilon_{tki} \quad (5.1)$$

$$\beta_{tk0} = \beta_{t-1,k0} + \beta_{k0} + \eta_{tk0} \quad (5.2)$$

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}, \quad j = 1, \dots, 4,$$

with the error terms ϵ_{tki} and η_{tkj} satisfying the assumptions (2.1), (2.4) and (2.5). Notice that the model assumed for the intercept term is the local approximation to a linear trend defined under case (d) of Section (2.1). The model assumed for the other coefficients is the random walk model defined under case (b).

The regression defined by (5.1) forms the basis for the construction of an HPI adjusted for changes in quality. By fixing the values of the QMV's at their average population values which are constant over time, (the values of these variables are adjusted approximately every five years), average sale prices can be computed using (5.1) and these averages are comparable between months since they refer to homes of similar qualities.

Pfeffermann, Burck and Ben-Tuvia discuss the considerations in selecting the model defined by (5.2) for the regression coefficients. They show empirical results which validate the fitness of the model. However, the results of that study were obtained by fitting the model to each cell separately, that is, without accounting for the cross-sectional relationships of the regression coefficients. This aspect of the model is explored in the present study. Another major purpose of the empirical study is to illustrate the performance of the modifications proposed in Section 4 to protect against model breakdowns.

5.2 Estimation of the Model

The model defined by (5.1) and (5.2) can be put in a state-space form similar to (2.7) and (2.8). In fact, the vectors α_t and the matrices Z_t , T and G assume, in this case, simple structures, since for $j = 1, \dots, 4$, $\beta_{kj} \equiv 0$ (see case (b) of Section 2.1). Thus, $\alpha'_{tk} = (\beta_{tk0}, \beta_{k0}, \beta_{tk1}, \dots, \beta_{tk4})$, $Z_{tk} = [1_{ntk}, 0_{ntk}, X_{tk}^{(1)}, \dots, X_{tk}^{(4)}]$, $\tilde{T} = [\epsilon_1, \epsilon_1 + \epsilon_2, \epsilon_3, \dots, \epsilon_6]$, a 6×6 matrix with ϵ_j having a one in position j and zeroes elsewhere and $\tilde{G} = [\epsilon_1, \epsilon_3, \dots, \epsilon_6]$ which is 6×5 . The matrix Δ is defined as in (2.5). The vector α_t and the matrices Z_t , T , G and Λ are obtained from the vectors $\{\alpha_{tk}\}$ and the matrices $\{Z_{tk}\}$, \tilde{T} , \tilde{G} and Δ in the same way as in (2.7) and (2.8).

Having set the model in a state-space form we next attempted to estimate the unknown variances and covariances using the method of scoring algorithm described in Section 3.2. As it turned out, however, the computer time needed for convergence was way beyond the capacity of the IBM 1481 mainframe used for this study. Notice that the number of unknown parameters of the combined state-space model is $\dim(\lambda) = 25$ whereas the dimension of the

state vectors and hence the dimension of the corresponding V-C matrices is $\dim(\alpha_t) = 30$. The total number of observations per month ranges from 55 to 353. The computer program written for this study uses numerical derivatives so that each iteration of the method of scoring requires a separate sweep through all the data with each sweep involving $[\dim(\lambda) + 1]$ computations of the state vector $\hat{\alpha}_t$ and the V-C matrix P_t (equation 3.1) at each point in time. These computations are needed in order to evaluate the log likelihood functions and hence the corresponding derivatives. It is clear therefore that the computational costs increase with the length of the series, the number of observations, the size of the state vector and the number of unknown parameters.

In order to deal with this problem we estimated the variance σ_k^2 (equation 2.1) and the matrix Δ (equation 2.5) separately for each of the five apartment sizes using the time series of observations corresponding to each size and then estimated the correlations ρ_j (equation 2.6) by a crude, grid search procedure. We found that setting $\rho_j = 1/2$ for every j gives satisfactory results both in terms of the behaviour of the innovations (the one step ahead prediction errors) and in terms of the smoothness of the regression coefficients corresponding to apartments of size one and five where the monthly sample sizes are very small. Notice that by estimating the variances and covariances defining the time series relationships of the regression coefficients separately for each size, one is more flexible in terms of the model assumptions although there is some loss of efficiency if the variances and covariances are indeed the same across the different sizes.

5.3 Results

Pfeffermann, Burck and Ben-Tuvia (1989) illustrate the adequacy of the time series models fitted to the various apartment sizes. As mentioned earlier, our purpose in this study is to compare the results obtained with and without the accounting for the cross-sectional correlations and to illustrate the performance of the modifications (4.1) in protecting against model breakdowns.

In order to sharpen the comparisons as much as possible, we deliberately inflated the Y -values by 5 percent in each of the following four months: October 1987, November 1988, January 1989 and May 1989. Thus all the Y -values of all the apartment sizes corresponding to the months October 1987 – October 1988 were inflated by 5 percent, the Y -values corresponding to November 1988 – December 1988 were inflated by 10.25 percent (5 percent on top of the previous 5 percent) and so forth. These kinds of model breakdowns (although obviously not in such magnitudes) may result from intentional devaluations of the currency and are of main concern when modeling sale prices. See Pfeffermann, Burck and Ben-Tuvia for further discussion. Similar model breakdowns may occur, for example, with series of unemployment rates in periods of abrupt economic recessions.

Table 1 shows the average mean squared errors (AMSE) of the model residuals $\hat{\epsilon}_{t ki} = (Y_{t ki} - \hat{\beta}_{t k 0} - \sum_{j=1}^4 X_{t ki}^{(j)} \hat{\beta}_{t kj})$ and the model innovations $e_{t ki} = [Y_{t ki} - (\hat{\beta}_{t-1, k 0} + \beta_{k 0}) - \sum_{j=1}^4 X_{t ki}^{(j)} \hat{\beta}_{t-1, kj}]$ (see equations 5.1 and 5.2), separately for each of the five apartment sizes. The AMSE's were computed as $AMSE_k(\epsilon) = 1/N \sum_{t=1}^N (1/n_t \sum_{i=1}^{n_t} \hat{\epsilon}_{t ki}^2)$; $AMSE_k(e) = 1/N \sum_{t=1}^N (1/n_t \sum_{i=1}^{n_t} e_{t ki}^2)$ where $t = 1, \dots, N$ indexes the months of July 1987 – November 1989. We distinguish between four different estimators of the regression coefficients as defined by whether the model accounts for the cross-sectional correlations ($\rho_j \equiv 1/2$), ($\rho_j \equiv 0$) and by whether or not the estimators are modified to protect against the model breakdowns (abbreviated as "Rob. Inc." and "No Rob." in the table). The modifications were carried out by augmenting the observation equation of each month by three linear constraints of the form 4.2. These constraints forced the aggregate means of the fitted values in each of the three

Table 1

Average Mean Squared Errors of Residuals and Innovations With and Without the Accounting for Cross-sectional Correlations and the Inclusion of the Robustness Modifications, by Size

Apt. Size	Mean Squared Errors of Innovations				Mean Squared Errors of Residuals			
	$\rho \equiv \frac{1}{2}$		$\rho \equiv 0$		$\rho \equiv \frac{1}{2}$		$\rho \equiv 0$	
	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.	Rob. Inc.	No Rob.
1	.141	.134	.176	.218	.021	.027	.056	.092
2	.070	.090	.084	.123	.021	.039	.023	.070
3	.065	.090	.070	.197	.017	.042	.019	.143
4	.067	.123	.072	.198	.019	.066	.021	.141
5	.067	.114	.077	.193	.023	.033	.065	.106

districts to coincide with the corresponding means of the observed values. When incorporating the constraints, the model was fitted using the amended Kalman filter as defined by the equations (4.3) and (4.4).

In order to illustrate the performance of the four sets of regression estimators in the various months and in particular, in and around the months where we inflated the data, we plotted the monthly MSE's of the innovations and residuals as obtained for 3 and 5 room apartments. The plots are shown in Figures 1 to 4. Notice that the values of Table 1 for 3 and 5 room apartments are correspondingly the averages of the values shown in the four figures.

The main conclusions from the table and the graphs are as follows:

Accounting for the cross-sectional correlations and including the linear constraints to protect against the model breakdowns yields better results than in the other cases considered. This outcome is most prominent in the cells of 1 and 5 room apartments where the sample sizes in each month are very small. In the other three cells, there are only small differences between the case ($\rho \equiv \frac{1}{2}$, Rob. Inc.) and the case ($\rho \equiv 0$, Rob. Inc.) which could be expected since as the number of observations in each month increases, there is less borrowing of information from neighbouring cells (small areas in the more general context). The situation is different, however, when the linear constraints are removed. Accounting for the cross-sectional correlations yields in this case much better results than when not accounting for them and this is true for all the apartment sizes. Thus, by borrowing information from one cell to the other, the estimators of the regression coefficients adapt themselves much more rapidly to the sudden drifts in the data as seen also more directly in the figures [The four peaks in each graph are in the months where the data were inflated and as can be seen, the graphs corresponding to the case ($\rho \equiv \frac{1}{2}$, No Rob.) return to their normal level of the months before the inflation much faster than the graphs representing the case ($\rho \equiv 0$, No Rob.)]

Another interesting comparison is between the case where the linear constraints are included and the case where they are not. Clearly, the inclusion of the constraints improves the results substantially when accounting for the serial correlations and the improvements are even more prominent when the serial correlations are set to zero. It is interesting to compare in this context the figures exhibiting the monthly MSE's of the innovations with the figures exhibiting the monthly MSE's of the residuals. In the four months where we inflated the data the MSE's of the innovations are high which is obvious since the innovations are the differences between the observations and their predictors from previous months. Still, when the linear constraints are included, the MSE's return to their normal level right after the months of inflation. As

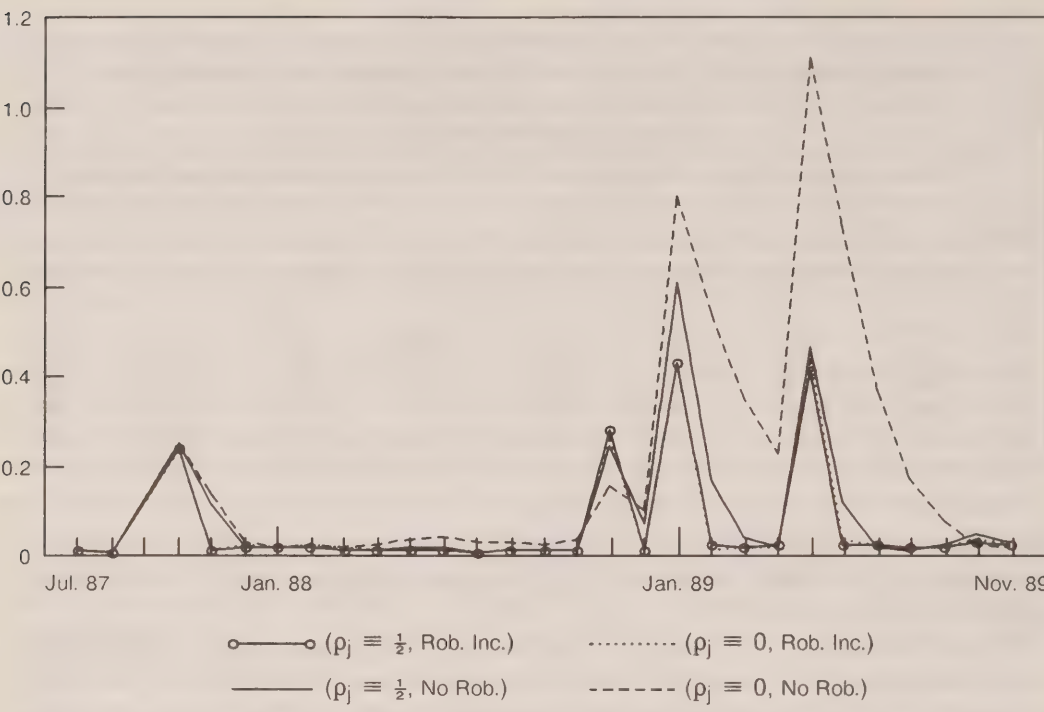


Figure 1 Monthly Mean Squared Errors of Innovations, 3 Room Apartments

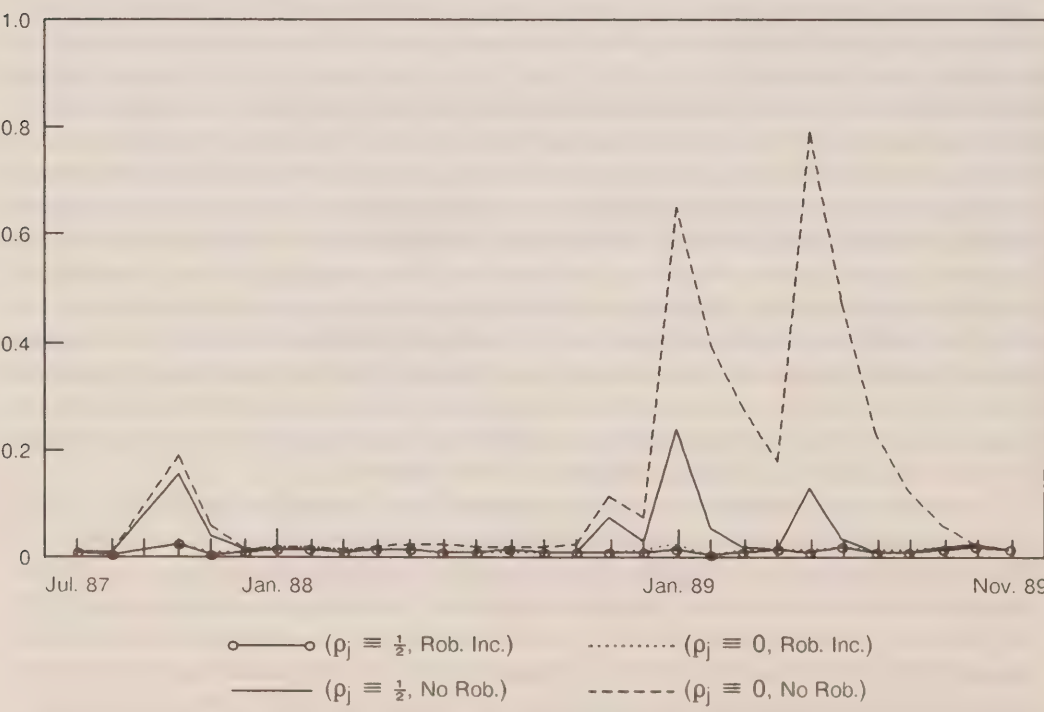


Figure 2 Monthly Mean Squared Errors of Residuals, 3 Room Apartments

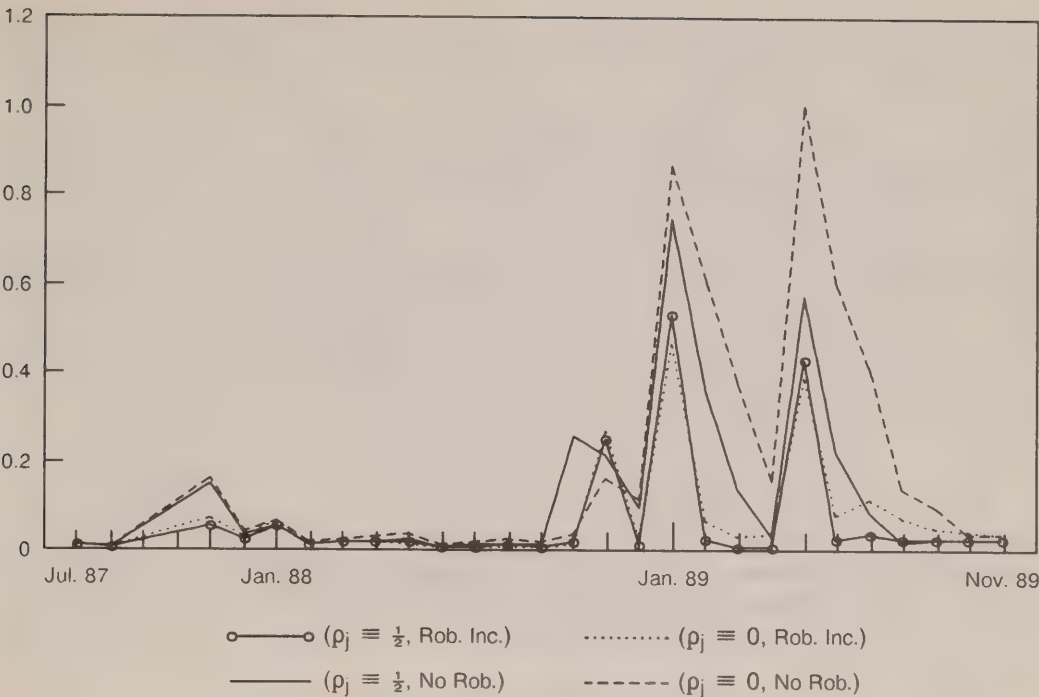


Figure 3 Monthly Mean Squared Errors of Innovations, 5 Room Apartments

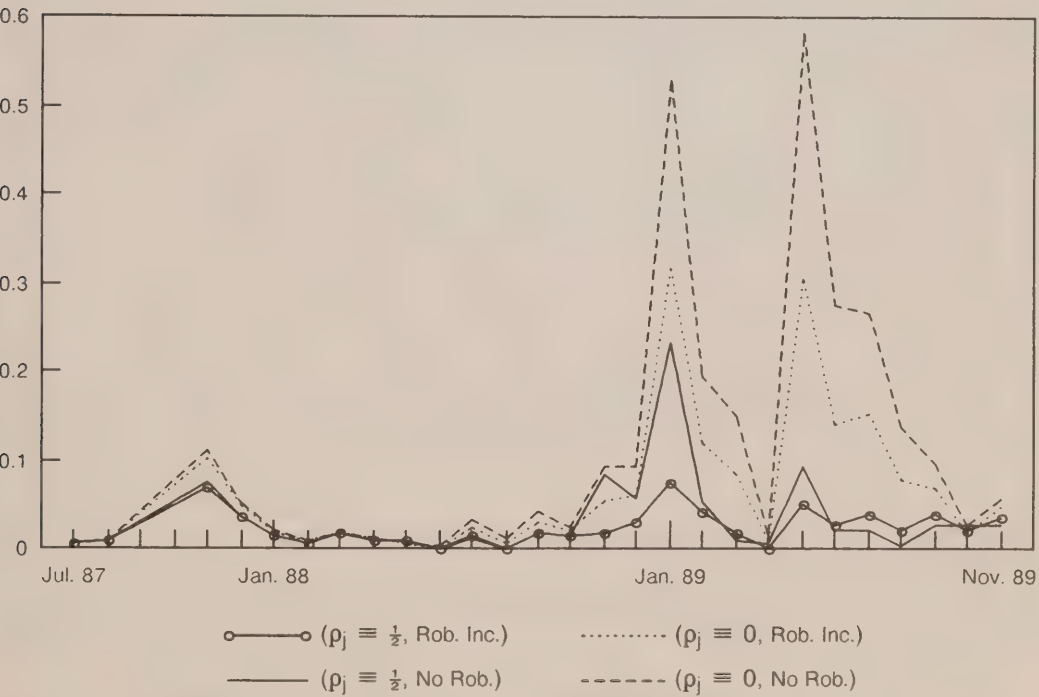


Figure 4 Monthly Mean Squared Errors of Residuals, 5 Room Apartments

for the residuals, once the linear constraints are included, there is practically no increase in the MSE values in the months of inflation in the case of 3 room apartments and, when accounting for the serial correlations, only a slight increase in the case of 5 room apartments. However, when ignoring the serial correlations, the residual MSE's for 5 room apartments are much larger in the months of inflation than in the other months even when imposing the constraints. This outcome has a simple explanation. The linear constraints are imposed on the aggregate means of the fitted values in each district but since the number of observations in 5 room apartments is a small fraction of the total number of observations, the constraints alone have a relatively small effect on the estimated regression coefficients in this cell. On the other hand, the constraints have a large effect on the estimated coefficients in the other cells so that when accounting for the cross-sectional correlations, the estimators corresponding to 5 room apartments are also modified since they are correlated with the other coefficients.

The way by which the linear constraints protect against sudden drifts in the data is illuminated in Figure 5 where we plotted the monthly intercept estimates for 3 room apartments.

As can be seen, with the linear constraints included, the intercept adapts itself to the new level of the data in the same month that the inflation occurs. Without the inclusion of the constraints, the adaption to the new level of the data takes several months. The plot of the monthly intercept estimates of 5 room apartments does not have this nice pattern since with the small sample sizes observed each month, the effect of the inflation is to alter also the other regression coefficients.

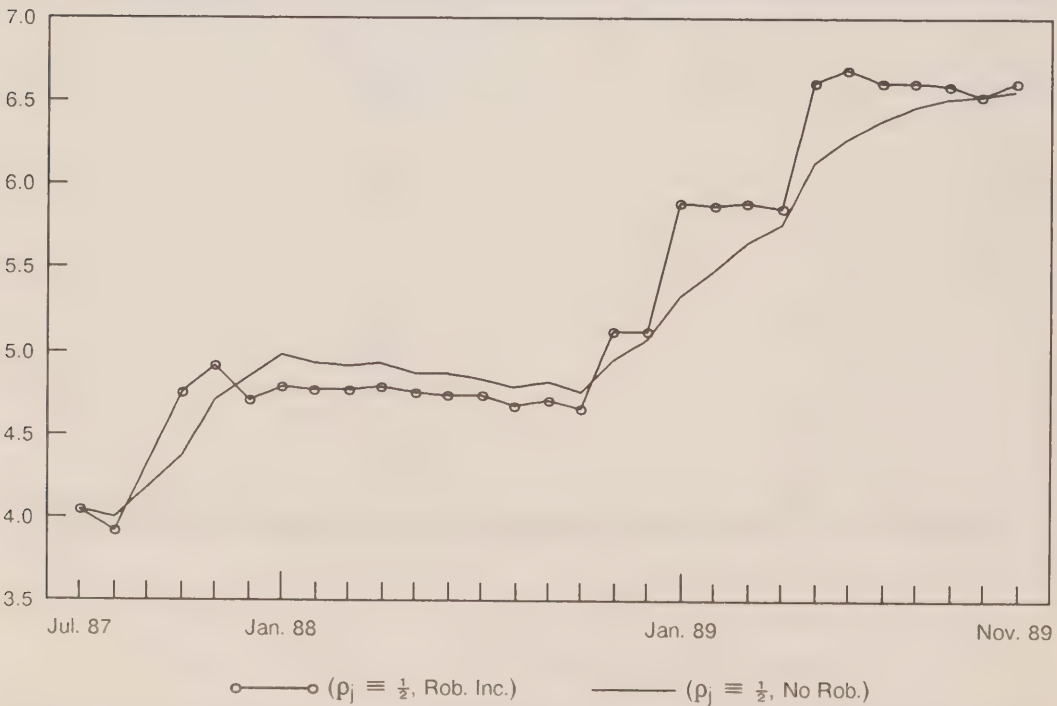


Figure 5 Monthly Estimates of Intercept, 3 Room Apartments

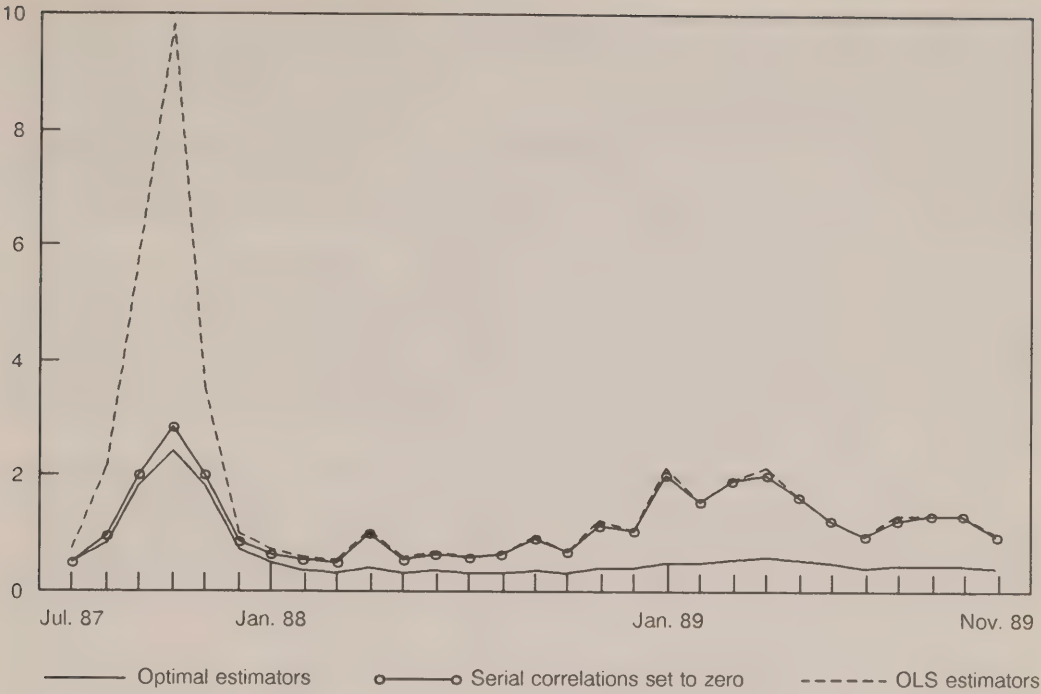


Figure 6 Variances of Estimators of Cell Means ($\times 10^4$), 3 Room Apartments

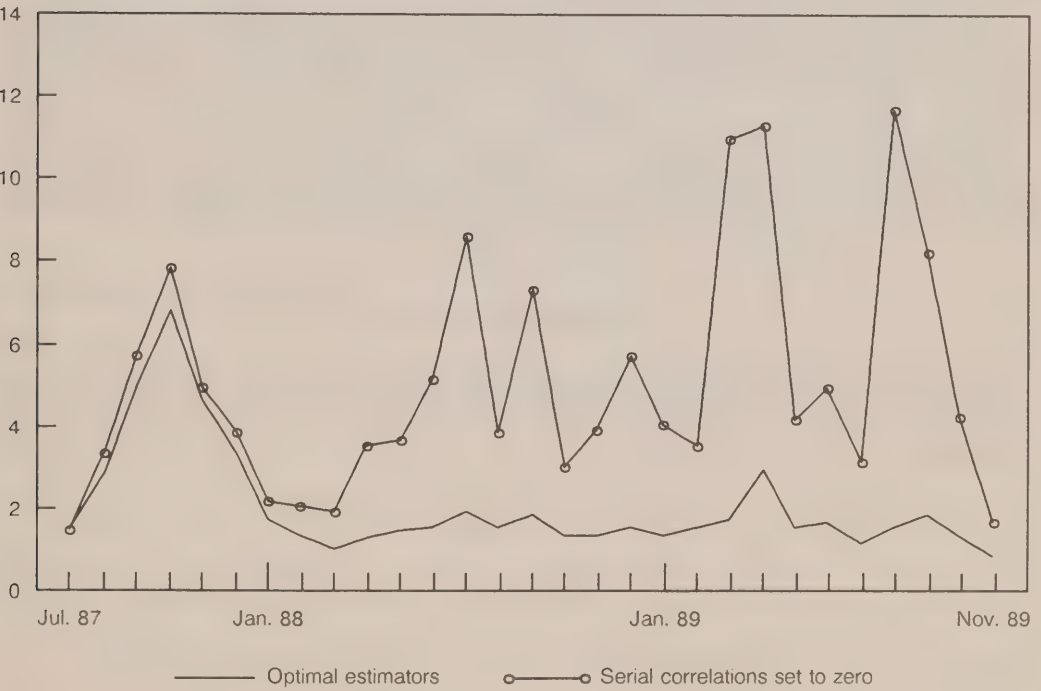


Figure 7 Variances of Estimators of Cell Means ($\times 10^4$), 5 Room Apartments

Our discussion so far centered on the empirical distribution of the model residuals and innovations. A major application of small area estimation is the prediction of the small area means (equation 2.2). Clearly, when a model yields residuals with well behaved properties it can also be expected to yield good estimators for the population means. Nevertheless, it is interesting to compare the theoretical variances of the small area means estimators as obtained with and without the accounting for the cross-sectional correlations, under the model which accounts for these correlations with $\rho_j \equiv 1/2$. This comparison permits the assessment of the loss in efficiency when the serial correlations are ignored.

Figures 6 and 7 show the monthly variances of the cell mean estimators as obtained for 3 and 5 room apartments. (The variances have been multiplied by 10^4 .) The figure for 3 room apartments also contains the variances of the ordinary least squares (OLS) estimators of the population means, that is, the variances of the estimators when estimating the regression coefficients in each month by OLS. These estimators are not operational in the case of 5 room apartments because of the very small monthly sample sizes.

The important conclusion drawn from the two figures is that by accounting for the cross-sectional correlations the variances of the resulting estimators can be reduced quite substantially, depending on the sample sizes. This is obviously the case in the case of 5 room apartments but is also true for 3 room apartments despite the fact that the sample sizes in these cells are relatively very large. The large sample sizes ordinarily obtained for 3 room apartments make the OLS estimators quite comparable to the estimators obtained when ignoring the cross-correlations in the estimation of the population means. Notice however the big gap between the variance of the OLS estimator and the variance of the other two estimators in October 1987. In this month there were only 10 observations of 3 room apartments and it is here where the use of the past data has its main impact even when ignoring the cross-sectional correlations. (The number of observations for 3 room apartments in November 1987 is 28; in all the other months there are at least 46 observations.)

Another important outcome arising from the two figures is the much greater stability of the variances of the optimal estimators under the model as compared to the variances of the estimators which ignore the cross-sectional correlations. Notice in this respect that the differences in the variances from one month to the other depend not only on the sample sizes in each month but also on the values of the explanatory variables (the design matrix) and the amount of past data observed. Still, it is the sample sizes which mostly explains the differences in the variances of the estimators particularly towards the end of the series.

ACKNOWLEDGEMENT

This article was written while the first author was on sabbatical leave at Statistics Canada under its Research Fellowship program. The authors would like to thank a referee for helpful comments.

APPENDIX

a) Derivation of Equation (2.12)

When $x_{tki} = x_{tk}$, $\hat{\Theta}_{tk} = x'_{tk} \hat{\Theta}_{tk} = z'_{ik} \hat{\Theta}_{tk}$ so that $\hat{\Theta}_t = (\hat{\Theta}_{t1}, \dots, \hat{\Theta}_{tK})' = Z_t \hat{\Theta}_t$. Also, for the random walk model the matrix T is the identity matrix and by equation (3.1)

$$Z_t \hat{\Theta}_t = Z_t \hat{\Theta}_{t-1} + (Z_t P_{t|t-1} Z_t') F_t^{-1} (Y_t - Z_t \hat{\Theta}_{t-1}) = (I - \sum_t F_t^{-1}) Y_t + \sum_t F_t^{-1} Z_t \hat{\Theta}_{t-1} \quad (A1)$$

since $F_t = (Z_t P_{t|t-1} Z_t' + \Sigma_t)$. Suppose for convenience that $k = 1$ and define

$$F_t = \begin{bmatrix} f_{11}, f_1' \\ f_1, F_{22} \end{bmatrix} \quad \text{and} \quad H_t = F_t^{-1} = \begin{bmatrix} h_{11}, h_1' \\ h_1, H_{22} \end{bmatrix} \quad \text{where } f_{11} \text{ and } h_{11}$$

are scalars, f_1' and h_1' are $[1 \times (K - 1)]$ and F_{22} and H_{22} are $[(K - 1) \times (K - 1)]$. Using this notation, it follows from (A1) that

$$\hat{\Theta}_{t1} = \left(1 - \frac{\sigma_1^2}{n_{t1}} h_{11}\right) \bar{Y}_{t1} + \frac{\sigma_1^2}{n_{t1}} h_{11} (x'_{t1} \hat{\Theta}_{t-1,1}) - \frac{\sigma_1^2}{n_{t1}} \sum_{k=2}^K h_{11} \frac{h_{1k}}{h_{11}} \bar{e}_{tk}. \quad (A2)$$

Let $\gamma_1' = (\gamma_{12}, \dots, \gamma_{1K}) = f_1' F_{22}^{-1}$ defines the partial regression coefficients in the regression of \bar{e}_{t1} on $(\bar{e}_{t2}, \dots, \bar{e}_{tK})$ and $v_1^2 = (f_{11} - f_1' F_{22}^{-1} f_1)$ define the residual variance in the regression.

Equation (2.12) follows directly from (A2) since

$$f_1' F_{22}^{-1} = -\frac{1}{h_{11}} h_1'; \quad (f_{11} - f_1' F_{22}^{-1} f_1)^{-1} = h_{11} \quad (A3)$$

by well known properties of the inverse of a partitioned matrix.

b) Derivation of Equation (4.4)

By (4.3),

$$\hat{\Theta}_t^{(A)} = (I - K_t^{(P)} Z_t^{(A)}) T \hat{\Theta}_{t-1}^{(A)} + K_t^{(P)} Y_t^{(A)}. \quad (A4)$$

Hence,

$$\hat{\Theta}_t^{(A)} - \alpha_t = (I - K_t^{(P)} Z_t^{(A)}) (T \hat{\Theta}_{t-1}^{(A)} - \alpha_t) + K_t^{(P)} (Y_t^{(A)} - Z_t^{(A)} \alpha_t). \quad (A5)$$

The prediction errors $(T \hat{\Theta}_{t-1}^{(A)} - \alpha_t)$ are independent of the residuals $(Y_t^{(A)} - Z_t^{(A)} \alpha_t)$ and so,

$$P_t^{(A)} = E[(\hat{\Theta}_t^{(A)} - \alpha_t)(\hat{\Theta}_t^{(A)} - \alpha_t)'] = Q_t P_{t|t-1}^{(A)} Q_t' + K_t^{(P)} \Sigma_t^{(A)} K_t^{(P)'} \quad (A6)$$

where we denote for convenience $Q_t = (I - K_t^{(P)} Z_t^{(A)})$.

By definition of the matrix $F_t^{(P)}$ (see below 4.3), equation (A6) can be written in the form

$$P_t^{(A)} = Q_t P_{t|t-1}^{(A)} - P_{t|t-1}^{(A)} Z_t^{(A)'} K_t^{(P)'} + K_t^{(P)} F_t^{(P)} K_t^{(P)'} + K_t^{(P)} (\Sigma_t^{(A)} - \Sigma_t^{(P)}) K_t^{(P)'} \quad (A7)$$

which implies the relationship (4.4) by straightforward algebra.

REFERENCES

- ANDERSON, B.O.D., and MOORE, J.B. (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.
- ANSLEY, C.F., and KOHN, R. (1986). Prediction mean squared error for State Space models with estimated parameters. *Biometrika*, 73, 467-473.
- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, Vol. 6, (Eds.), P.R. Krishnaiah and C.R. Rao, Amsterdam: Elsevier Science, 187-211.
- CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Analysis of Data in Time*. Ottawa: Statistics Canada (to appear).
- COOLEY, T.F., and PRESCOTT, E.C. (1976). Estimation in the presence of stochastic parameter variation. *Econometrica*, 44, 167-184.
- DIELMAN, T.E. (1983). Pooled cross-sectional and time series data: A survey of current statistical methodology. *The American Statistician*, 37, 111-122.
- HAMILTON, J.D. (1986). A standard error for the estimated state vector of a State-Space model. *Journal of Econometrics*, 33, 388-397.
- HARVEY, A.C. (1981). *Time Series Models*. Deddington, Oxford: Philip Allan.
- HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- KITAGAWA, G., and GERSCH, W. (1984). A smoothness priors State-Space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79, 378-389.
- JOHNSON, L.W. (1977). Stochastic parameter regressions: An annotated bibliography. *International Statistical Review*, 45, 257-272.
- JOHNSON, L.W. (1980). Stochastic parameter regression: An additional annotated bibliography. *International Statistical Review*, 48, 95-102.
- LaMOTTE, L.R., and McWHORTER, A. (1977). Estimation, testing and forecasting with random coefficient regression models. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 814-817.
- MADDALA, G.S. (1977). *Econometrics*. Kogakusta: McGraw-Hill.
- MEINHOLD, R.J., and SINGPURWALLA, N.D. (1983). Understanding the Kalman filter. *The American Statistician*, 37, 123-127.

- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 163-175.
- PFEFFERMANN, D., and BARNARD, C. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* (to appear).
- PFEFFERMANN, D., BURCK, L., and BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Analysis of Data in Time*. Ottawa: Statistics Canada.
- PFEFFERMANN, D., and SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.
- ROSENBERG, B. (1973a). The analysis of cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, 2, 399-428.
- ROSENBERG, B. (1973b). A survey of stochastic parameter regression. *Annals of Economic and Social Measurement*, 2, 381-397.
- SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHWEPPE, F. (1965). Evaluation of likelihood functions for gaussian signals. *IEEE Transactions on Information Theory*, 11, 61-70.
- SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (Ed.) N.K. Nawboodivi, New York: Academic Press, 201-216.
- SWAMY, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.
- TILLER, R. (1989). A Kalman filter approach to labour force estimation using survey data. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association* (to appear).
- WATSON, M.W., and ENGLE, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23, 385-400.

A Method for the Analysis of Seasonal ARIMA Models

DAVID A. BINDER and J. PETER DICK¹

ABSTRACT

A commonly used model for the analysis of time series models is the seasonal ARIMA model. However, the survey errors of the input data are usually ignored in the analysis. We show, through the use of state-space models with partially improper initial conditions, how to estimate the unknown parameters of this model using maximum likelihood methods. As well, the survey estimates can be smoothed using an empirical Bayes framework and model validation can be performed. We apply these techniques to an unemployment series from the Labour Force Survey.

KEY WORDS: Kalman filter; Partial likelihood; Data smoothing.

1. INTRODUCTION

It is common practice to analyze data from surveys where similar data items are collected on repeated occasions, using time series analysis methods. Most standard methods for these analyses assume the data are either observed without error or have independent measurement errors. However, in the analysis of repeated survey data, when there are overlapping sampling units between occasions, the survey errors can be correlated over time.

A commonly used model in the analysis of time series is the seasonal integrated autoregressive-moving average (ARIMA) regression model, which we discuss in this paper. We show how to incorporate the (possibly correlated) survey errors into the analysis. In particular, we consider the case where the survey (design) error can be assumed to be an ARMA process up to a multiplicative constant.

When such a model for the behaviour of the population characteristics is assumed, the minimum mean squared error, or, equivalently, the Bayes linear estimator for the characteristic at a point in time can be derived. This estimator incorporates the model structure which the classical estimators, such as the minimum variance linear unbiased estimators, ignore. When the model parameters are estimated from the survey data, the estimators are empirical Bayes.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Jones (1980), Rao, Srinath and Quenneville (1989) and others considered the implications of certain stochastic models for the population means over time. Hausman and Watson (1985) incorporate a measurement error model into the standard seasonal adjustment process. Miazaki (1985) assumed that the survey error could be modelled with a pure moving average process. In Binder and Dick (1989), these results were generalized using state space models and Kalman filters. In this paper, we extend the framework to include the model where differencing of the original series of the population means yields an ARMA model. We use the modified Kalman filter approach given by Kohn and Ansley (1986). To estimate the unknown parameters, we maximize the marginal likelihood function using the method of scoring. This approach can also handle missing data routinely. We also show how the survey estimates can be smoothed to incorporate the model features using empirical Bayes methods. Confidence intervals for these

¹ D.A. Binder, Business Survey Methods Division and J.P. Dick, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

smoothed values are also given, using the method described by Ansley and Kohn (1986). Bell and Hillmer (1987) used a similar model but their initial conditions do not extend easily to include regression terms or missing values (while preserving the marginal likelihood approach).

An example of this model is described in Section 5 using unemployment data from the Canadian Labour Force Survey. This example shows the implications on the estimates of the model parameters when the survey errors are taken into account. We derive a smoothed estimate of the underlying process under the model assumptions. Recursive residuals are produced and validation techniques are used to evaluate the various models.

2. THE MODEL

Suppose we have a series of point estimates from a repeated survey of a population characteristic, given by y_1, y_2, \dots, y_T . We assume that y_t can be decomposed into three components, so that

$$y_t = x_t' \gamma + \theta_t + e_t, \quad (2.1)$$

where $x_t' \gamma$ is a deterministic regression term, θ_t is a population parameter following a time series model, and e_t is the survey error, assumed to have zero expectation.

We first describe an integrated seasonal autoregressive-moving average model for $\{\theta_t\}$. We let B be the backshift operator; $\nabla = 1 - B$ and $\nabla_s = 1 - B^s$, where s is the seasonal period. We define the following polynomial functions:

$$\lambda(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \dots - \lambda_P B^P,$$

$$\alpha(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \dots - \alpha_p B^p,$$

$$v(B) = 1 - v_1 B - v_2 B^2 - \dots - v_Q B^Q,$$

and

$$\beta(B) = 1 - \beta_1 B - \beta_2 B^2 - \dots - \beta_q B^q.$$

The seasonal ARIMA $(p, d, q)(P, D, Q)_s$ model for $\{\theta_t\}$ is given by

$$\lambda(B^s) \alpha(B) \nabla^d \nabla_s^D \theta_t = v(B^s) \beta(B) \epsilon_t, \quad (2.2)$$

where the ϵ_t 's are independent $N(0, \sigma^2)$. We define $a(B) = \lambda(B^s) \alpha(B)$, a $(p + sP)$ -degree polynomial; $\Delta(B) = \nabla^d \nabla_s^D$, a $(d + sD)$ -degree polynomial; $b(B) = v(B^s) \beta(B)$, a $(q + sQ)$ -degree polynomial; $A(B) = a(B) \Delta(B)$, a $(p + d + sP + sD)$ -degree polynomial; $u_t = \Delta(B) \theta_t$, an ARMA $(p + sP, q + sQ)$ process. Therefore, alternative representations of (2.2) are

$$a(B) \Delta(B) \theta_t = b(B) \epsilon_t, \quad (2.3)$$

$$A(B) \theta_t = b(B) \epsilon_t, \quad (2.4)$$

and

$$a(B) u_t = b(B) \epsilon_t. \quad (2.5)$$

We now consider the survey errors $\{e_t\}$ of expression (2.1). It will be assumed that the sample sizes of the repeated survey are sufficiently large that the errors for the survey estimates can be approximated by a multivariate normal distribution. In the simplest case, where the surveys are non-overlapping and the sampling fractions are small, the e_t 's can be assumed to be independent. In a rotating panel survey, the survey errors are usually correlated. In this case, since the correlations between survey occasions are zero after panels have been rotated out, a pure moving average process can be used to describe the survey error process.

Alternatively, if a random sample of units are replaced on each survey occasion, a pure autoregressive process may best describe the process. More complicated models are also possible. For example, in a two-stage design, some of the first stage units may be replaced randomly on each occasion and the second stage units may have a rotating panel design. This might be approximated by an autoregressive-moving average process, as suggested by Scott, Smith and Jones (1977).

In this paper, we assume that the survey error process is given by

$$e_t = k_t \omega_t, \tag{2.6}$$

where $\{\omega_t\}$ is an ARMA (m,n) process, given by

$$\phi(B)\omega_t = \psi(B)\eta_t \tag{2.7}$$

and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_m B^m,$$

and

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_n B^n.$$

The η_t 's are independent $N(0,\tau^2)$. The factor k_t has been included in (2.6) to allow for non-homogeneous variances when the autocorrelation function is homogeneous in time.

In the model just described we assume that τ^2 , the k_t 's and the coefficients of $\phi(B)$ and of $\psi(B)$ can be estimated directly from the survey data, using design-based methods. However, in general, the other parameters are unknown. This includes γ , σ^2 , and the coefficients of $\lambda(B)$, $\alpha(B)$, $v(B)$ and of $\beta(B)$. The x_t 's in the regression term are assumed known.

3. STATE SPACE FORMULATION OF THE MODEL

3.1 General Formulation

The model described in Section 2 can be formulated as a state space model with partially improper priors. This has a number of advantages. It permits, through use of a modified Kalman filter, calculation of a marginal likelihood function, which can be maximized to estimate unknown parameters. It also accommodates smoothing of the original survey estimates, by removing the estimates of survey error from the data.

In the state space model, two processes occur simultaneously. The first process, the observation system, details how the observations depend on the current state of the process parameters. The second process, the transition system, details how the parameters evolve over time.

For the state space models we consider here, the observation equation is written as

$$y_t = h'_t z_t \quad (3.1a)$$

and the transition equation is

$$z_t = Fz_{t-1} + G\xi_t, \quad (3.1b)$$

where z_t is an $(r \times 1)$ state vector and h_t is a fixed $(r \times 1)$ vector. In the transition equation, F is a fixed $(r \times r)$ transition matrix, G is a fixed $(r \times m)$ matrix and the ξ_t 's are independent normal vectors with mean zero and covariance U .

The final requirement to complete the specification of the state space process is the initial conditions for z_0 . In this paper, we shall use the improper prior formulation given in Kohn and Ansley (1986). In general, we assume that z_0 has a partially diffuse r -variate normal distribution with mean $m(0 | 0) = 0$ and covariance matrix $V(0 | 0)$, where

$$V(0 | 0) = \kappa V_1(0 | 0) + V_0(0 | 0) \quad (3.2)$$

for large κ . The matrix $V_1(0 | 0)$ specifies the diffuse part of the prior. We explain in Section 3.2 how to obtain $V_1(0 | 0)$ and $V_0(0 | 0)$ for our model.

We denote the conditional mean of z_t given the observations up to and including time t' by $m(t | t')$, and the conditional variance by $V(t | t')$, where

$$V(t | t') = \kappa V_1(t | t') + V_0(t | t'). \quad (3.3)$$

Recursive formulae for the cases where $t = t'$ and $t = t' + 1$ are given in Kohn and Ansley (1986). They refer to this as the modified Kalman filter.

Since the model for $\{y_t\}$ given by (2.1) contains survey errors $\{e_t\}$ an estimate of the components without survey error, given by

$$y_t \text{ (smoothed)} = x'_t \gamma + \theta_t \quad (3.4)$$

is often of interest. When the right hand side of (3.4) can be expressed as $g'_t z_t$, for some g'_t , then it is possible to obtain the conditional mean and variance of the linear combination $g'_t z_t$ given all the data, using the modified Kalman filter. To do this, the recursions are applied up to time t to obtain $m(t | t)$ and $V(t | t)$. Then the state vector z_t is augmented by the state $z_{t,r+1} = g'_t z_t$, and $m(t | t)$ and $V(t | t)$ are also appropriately augmented. The matrix F in (3.1b) is modified to add the equation $z_{t+1,r+1} = z_{t,r+1}$. After these modifications, the modified Kalman filter can be used as before, so that the last component of $m(T | T)$ gives the conditional expectation of $g'_t z_t$, given all the data, y_1, y_2, \dots, y_T . As well, the last diagonal component of $V(T | T)$ gives the conditional variance. This procedure can be generalized to include any number of smoothed estimates and their conditional covariances. In applications, space limitations on the computer might preclude computing the smoothed values for a large number of time points.

3.2 Model for θ

Harvey and Phillips (1979) described a method to put the ARIMA model (2.4) into the state space form given by (3.1). The dimension of z_t is $r = \max(p + d + sP + sD, q + sQ)$. By augmenting $A = (A_1, \dots, A_{p+d+sP+sD})$ or $b = (b_1, \dots, b_{q+sQ})$ with zeroes

to have dimension r , the ARIMA model may be written in the form given by (3.1), where $\mathbf{h}'_t = (1, 0, \dots, 0)$, $\mathbf{G}'_t = (1, -b_1, \dots, -b_{r-1})$ and

$$\mathbf{F} = \left[\begin{array}{c|c} A_1 & I_{r-1} \\ \vdots & \\ A_{r-1} & \\ \hline A_r & \mathbf{0}' \end{array} \right],$$

where I_{r-1} is the $(r-1)$ by $(r-1)$ identity matrix and $\mathbf{0}'$ is a row vector of zeroes.

In this formulation, the state vector $\mathbf{z}_t = (z_{1t}, \dots, z_{rt})'$ is defined as

$$\begin{aligned} z_{it} &= A_i \theta_{t-1} + A_{i+1} \theta_{t-2} + \dots + A_r \theta_{t-(r-i+1)} \\ &\quad - b_{i-1} \epsilon_t - b_i \epsilon_{t-1} - \dots - b_{r-1} \epsilon_{t-(r-i)}, \end{aligned} \quad (3.5)$$

for $i = 2, 3, \dots, r$ and $z_{1t} = \theta_t$.

To complete the specification for $\{\theta_t\}$, initial conditions for \mathbf{z}_0 are required. These are given in Ansley and Kohn (1985), a summary of which is provided here.

From expression (2.5), $\{u_t\}$ is an ARMA process. We define

$$\boldsymbol{\theta}_- = (\theta_0, \theta_{-1}, \dots, \theta_{-S})',$$

where $S = \max(0, p + sP + d + sD - 1)$. We let

$$\mathbf{u}_- = (u_0, u_{-1}, \dots, u_{-R})',$$

where $R = \max(0, p + sP - 1)$. Finally, we let

$$\mathbf{w}_- = (\theta_{-R-1}, \theta_{-R-2}, \dots, \theta_{-S})',$$

when $S > R$.

Now, \mathbf{u}_- is assumed to be a stationary ARMA process, so that its covariance matrix can be derived from expression (2.5). It is assumed that \mathbf{w}_- is $N(\mathbf{0}, \kappa \mathbf{I})$ and is independent of \mathbf{u}_- . Since $(\mathbf{u}_-, \mathbf{w}_-)'$ is a non-singular linear combination of $\boldsymbol{\theta}_-$, the covariance matrix for $\boldsymbol{\theta}_-$ can be derived. Using the form of expression (3.5) for \mathbf{z}_0 , the initial covariance matrix can be computed. Note that when both d and D are zero, so that no differencing takes place in the model, then \mathbf{w}_- is the null vector and we have $\mathbf{u}_- = \boldsymbol{\theta}_-$.

3.3 Model for the Observed Data

In Section 2 we assumed that $e_t = k_t \omega_t$, where ω_t is an ARMA(m, n) model. Therefore, from the discussion in Section 3.2, it is clear that e_t can be represented in state space form, with $\mathbf{h}_t = (k_t, 0, \dots, 0)'$, and $e_t = \mathbf{h}'_t \mathbf{z}_t$.

The regression component can be similarly represented by adding γ to the state vector and initially, assuming that γ has mean zero and covariance $\kappa \mathbf{I}$. Note that in the transition equation γ remains constant.

Since we can represent each of the components of y_t in expression (2.1) by a state space model, it is straightforward to combine the individual models into an overall model, by extending the state vector to include the state vectors from the individual components. The observation equation is then the sum of the three individual components.

4. ESTIMATION OF THE STATE SPACE MODEL

4.1 Estimation of the Parameters

The unknown parameters of this model are σ^2 , and the coefficients of $\lambda(B)$, $\alpha(B)$, $v(B)$ and $\beta(B)$. We transformed σ^2 to $\log(\sigma^2)$, in the numerical maximization procedure described below to avoid problems with negative parameter values. The model for the vector of observations $y = (y_1, y_2, \dots, y_T)'$ given in Section 3 is equivalent to

$$y = M\eta + \zeta, \quad (4.1)$$

where η is j -variate $N(0, \kappa I)$, ζ is T -variate $N(0, W)$, and M is some fixed $T \times j$ matrix. We note that η contains unknown constants including the regression coefficients; W is a function of the ARMA parameters; M is a function of the differencing structure.

Kohn and Ansley (1986) recommended maximizing the limit of $\kappa^{j/2}$ times the likelihood function for the data, as κ tends to infinity. It can be shown that this limit of the likelihood function is equivalent to the marginal likelihood function of $y - M\hat{\eta}$, where $\hat{\eta}$ is the maximum likelihood estimate of η when M and W are known. Tunnicliffe-Wilson (1989) has shown that the Jacobian of the transformation from the data y to $(\hat{\eta}, y - M\hat{\eta})$ does not depend on the model parameters of W whenever M is known. Ansley and Kohn (1985) have shown that M does not depend on the unknown parameters. By using the modified Kalman filter, the computations for the marginal likelihood function are more straightforward than the approach given by Tunnicliffe-Wilson.

The procedure we employed computes both the marginal likelihood function and its first derivatives with respect to the unknown parameters. This involves taking first derivatives of the initial conditions and of $m(t | t')$ and the components of $V(t | t')$ for $t = t'$ and $t = t' + 1$. All the computations were done using PROC IML in SAS.

The likelihood function was maximized using a modification of the method of scoring. This modification allowed for varying step sizes. On each iteration, the likelihood function was computed at the previous step size, as well as at this step size multiplied and divided by a predetermined constant. (We used 1.1 as the factor.) The next step size was to choose the point which maximized the likelihood function among the three points. Each time a check was made to determine whether the parameters were in range. This was done by checking for positive semi-definiteness of the initial covariance matrix of the state vector. If it was out of range, the step size was divided again by the constant and the procedure repeated.

To estimate the variance matrix for the estimated parameters, the inverse of the Fisher information matrix was used. This is readily computed since the first derivatives of the likelihood function are available.

4.2 Estimation of the Smoothed Values

Smoothed values as defined in (3.4) for the estimates can be obtained by zeroing out that component of the state vector which corresponds to the survey error. However, this still leaves open the question of how to estimate its variance. To derive the standard error of the smoothed

estimate it is necessary to account for the fact that the unknown parameters have been estimated from the data, particularly when the data series is short; see Jones (1979).

To obtain the variance of $\mathbf{g}'\mathbf{z}_t$, it is sufficient to derive the variance $\mathbf{z}_T - \hat{\mathbf{m}}(T | T)$, where $\hat{\mathbf{m}}(T | T)$ is the estimate of $\mathbf{m}(T | T)$ at the estimated parameter values. This is because the state vector has been augmented to include $\mathbf{g}'\mathbf{z}_t$. Now,

$$\begin{aligned}\mathbf{z}_T - \hat{\mathbf{m}}(T | T) &= [\mathbf{z}_T - \mathbf{m}(T | T)] \\ &+ [\mathbf{m}(T | T) - \hat{\mathbf{m}}(T | T)].\end{aligned}\quad (4.2)$$

The first component of the right hand side of (4.2) has conditional variance $V(T | T) = V_0(T | T)$, assuming that $V_1(T | T) = \mathbf{0}$. The second component of (4.2) represents a bias term and is independent of the first term, since it depends only on the data y . By taking a Taylor series expansion of the second term around the true parameter values and ignoring higher terms, we have the second component of (4.2) is

$$\mathbf{m}(T | T) - \hat{\mathbf{m}}(T | T) = \left[\frac{-\partial \hat{\mathbf{m}}(T | T)}{\partial \phi} \right]' (\hat{\phi} - \phi), \quad (4.3)$$

where ϕ is the vector of unknown parameters and $\hat{\phi}$ is its estimate. Therefore, the asymptotic variance of (4.2) is approximately

$$\begin{aligned}\text{Var}[\mathbf{z}_T - \hat{\mathbf{m}}(T | T)] &= V_0(T | T) \\ &+ \left[\frac{\partial \hat{\mathbf{m}}(T | T)}{\partial \phi} \right]' V_\phi \left[\frac{\partial \hat{\mathbf{m}}(T | T)}{\partial \phi} \right],\end{aligned}\quad (4.4)$$

where V_ϕ is the covariance matrix for the unknown parameters. Expression (4.4) is estimated by using the estimated parameter values. This is the same approach as that given by Ansley and Kohn (1986).

4.3 Generalized Recursive Residuals

As Harvey and Durbin (1986) pointed out, useful quantities for performing model diagnostics are the generalized recursive residuals. In terms of our state space model, this is the difference between the observation and the one-step ahead prediction from the Kalman filter. These can be used for all time points t where $V_1(t + 1 | t) = \mathbf{0}$. Under the model, these residuals are approximately independent normal. They can be standardized to have an estimated variance of unity under the model. Diagnostics similar to those used in classical regression models can then be performed.

5. ANALYSIS OF LABOUR FORCE DATA

5.1 Parameter Estimation

To demonstrate this procedure, we take data from the Canadian Labour Force Survey (LFS). The LFS is a monthly rotating panel survey with each panel containing one-sixth of the selected households. A panel will remain in the sample for six consecutive months while the primary sampling units will rotate out after approximately two years. The sample selection follows a stratified multi-stage design.

The data were the monthly number of unemployed as published from January 1977 to December 1986 for the province of Nova Scotia and for the subprovincial region within Nova Scotia corresponding to Cape Breton Island. This province was selected because the sampling errors are moderate compared to the larger provinces. Cape Breton Island was selected because its smaller sample size provides estimates with a larger relative variance. Graph 1a displays the logarithm of the Nova Scotia series and Graph 1b shows the similarly transformed Cape Breton Island series. We used the logarithms as our inputs.

Lee (1990) estimated the autocorrelations for the Nova Scotia survey error up to a lag of eleven. We derived the coefficients of the ARMA (m,n) survey error process given in (2.7) by matching these autocorrelations. A good fit was found using an ARMA (3,6) model. The resulting coefficients were:

$$\begin{aligned} \phi_1 &= 0.2575 & \psi_1 &= -0.1847 \\ \phi_2 &= -0.3580 & \psi_2 &= -0.5873 \\ \phi_3 &= -0.6041 & \psi_3 &= 0.3496 \\ & & \psi_4 &= 0.0647 \\ \tau^2 &= 0.7246 & \psi_5 &= 0.0982 \\ & & \psi_6 &= 0.0347. \end{aligned}$$

The k_i 's of (2.6) were the estimated standard errors of the estimates, derived by taking a Taylor series approximation for the logarithms.

A series of models were fitted to the Nova Scotia data with an assumption of no sampling error. The same models were then refitted, incorporating the model for the survey error process. In this case we could also compute smoothed values for the survey estimates and compare their standard errors with the standard errors of the original series.

The preliminary model selected for the Nova Scotia data, ignoring the sampling error, was a seasonal ARIMA $(1,1,0)(0,1,1)_{12}$. However the moving average term for the seasonal component was estimated to be one, so a deterministic regression term was used to account for the seasonality. The 12 regression variables included a linear term and a dummy variable for each of the first 11 months. The dummy variable for a reference month took the value 1 for the reference month, -1 for December and 0 for the other months. Note that an intercept term is not appropriate for this model because the first differences of the data are fitted.

Further analysis of this reduced model showed that the moving average seasonal component was not required in the model. The final model selected for the Nova Scotia data was an ARIMA $(1,1,0)$ with a deterministic regression component. This same model was then used for the Nova Scotia data with the survey error process incorporated. The same structural model was used for the Cape Breton Island series.

Table 1 displays the parameter estimates. The estimates that do not incorporate the survey error component are in the **Without Sampling Errors** columns. First, examining the models for Cape Breton Island shows that the regression estimates are similar, as would be expected. Note that the autoregressive estimates (AR) are also similar and that the **With Sample Error** model has reduced the estimated model variance substantially. The column headed **T-value** displays the estimated parameter divided by its standard error. Note that the t -values for the autoregressive parameter are substantially different (-0.68 vs -2.85). This would lead to

Table 1
Parameter Estimates – Unemployment Series 1977-1986

Parameter	Nova Scotia				Cape Breton Island			
	Without Sampling Error		With Sampling Error		Without Sampling Error		With Sampling Error	
	Estimate	T-value	Estimate	T-value	Estimate	T-value	Estimate	T-value
Alpha	-0.296	-3.23	0.862	2.08	-0.260	-2.85	-0.231	-0.68
Sigma	0.0597	-	0.0032	-	0.1049	-	0.0520	-
Trend	0.00427	1.01	0.00420	1.89	0.00607	0.79	0.00598	1.50
January	0.064	3.60	0.048	1.93	-0.007	-0.23	-0.003	-0.10
February	0.083	4.80	0.078	3.30	0.027	0.89	0.028	0.97
March	0.166	10.20	0.165	6.40	0.171	5.76	0.164	5.76
April	0.106	6.60	0.104	4.10	0.099	3.33	0.089	3.19
May	0.009	0.60	0.016	0.70	-0.008	-0.28	-0.007	-0.24
June	-0.101	-6.00	-0.088	-3.30	-0.029	-0.96	-0.033	-1.17
July	-0.016	-1.20	-0.014	-0.63	0.082	2.77	0.081	3.13
August	-0.058	-3.60	-0.062	-2.37	-0.011	-0.37	-0.009	-0.30
September	-0.106	-6.60	-0.105	-3.96	-0.104	-3.51	-0.098	-3.18
October	-0.081	-4.80	-0.071	-3.08	-0.084	-2.83	-0.069	-2.44
November	-0.026	-1.80	-0.029	-1.08	-0.063	-2.10	-0.074	-2.46

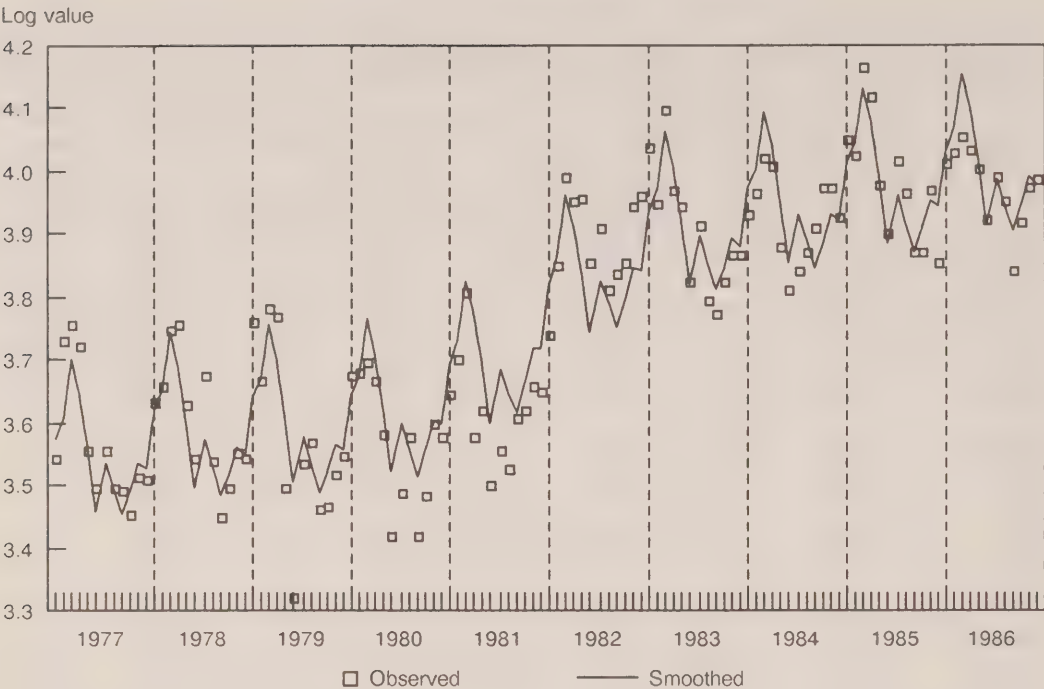
accepting a model for the Cape Breton Island data with only a deterministic regression term when the survey error process is incorporated into the model. However, if the survey error is ignored in the analysis, too much significance would be attached to the autoregressive parameter.

The results for the Nova Scotia models are also displayed on Table 1. Note that the reduction in the estimate of the model variance by incorporating the sampling error structure is much greater for the Nova Scotia series than was achieved for the Cape Breton data. An important result in the Nova Scotia models is the difference in the estimates for the autoregressive component. Both models show that the AR component is highly significant in each model. The **Without Sample Error** model gives an estimate of $\alpha = -0.296$; whereas the **With Sample Error** model gives an estimate of $\alpha = 0.862$. Clearly, the interpretations that would be associated with these two estimates are entirely different.

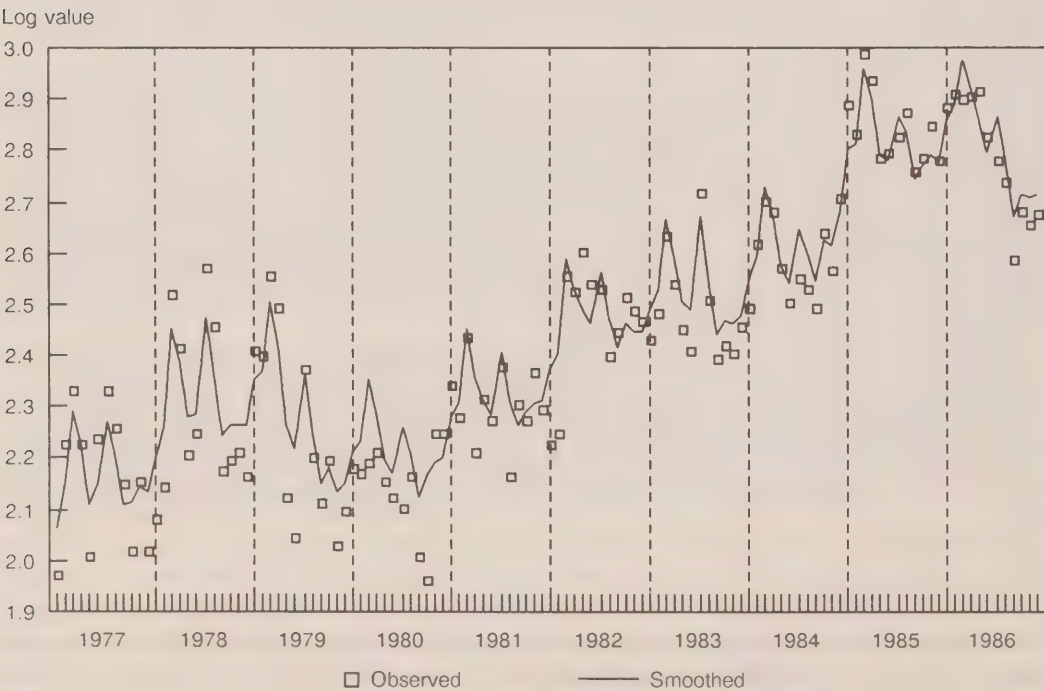
The smoothed estimates for the model incorporating sampling error are shown superimposed on the original data series in Graph 1a. Graph 1b shows the smoothed estimates for Cape Breton Island superimposed on the original series. The most notable item in these plots is the impact of the recession of 1981 on the smoothed estimates. Prior to the recession, the model tends to overestimate unemployment and after 1981 the model tends to underestimate the number of unemployed.

5.2 Model Validation

The plots of the generalized recursive residuals (described in Section 4.3) against the lagged generalized recursive residuals were produced for all the models. Graphs 2a and 2b show these plots for the two models for Nova Scotia. Note that Graph 2a shows less dispersion around the origin than Graph 2b, indicating a better fit when survey error is incorporated in the model.

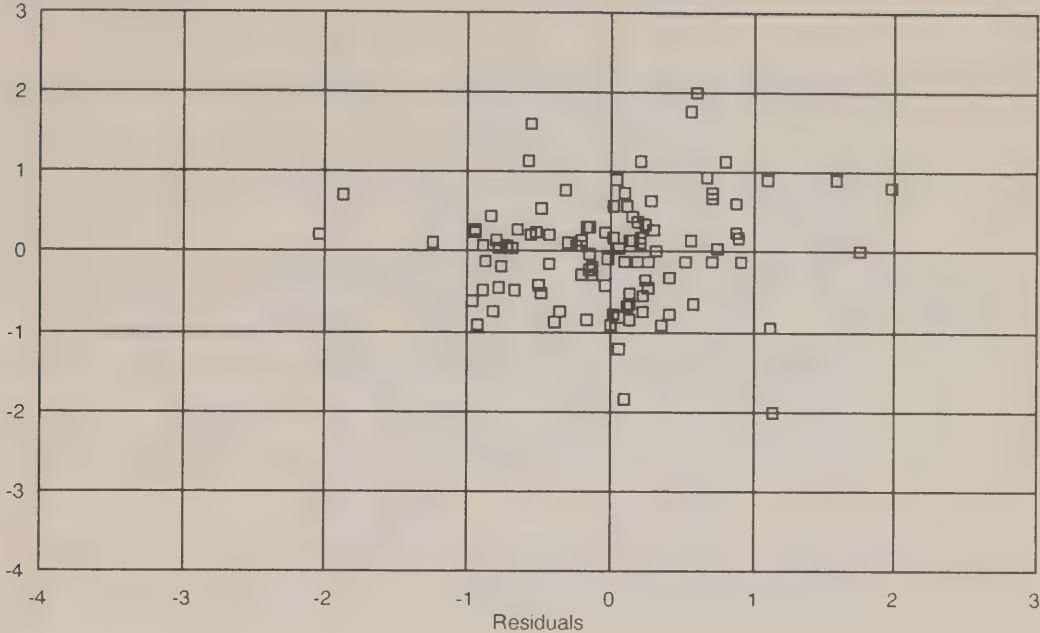


Graph 1a Nova Scotia Observed and Smoothed Values (Log Transform)



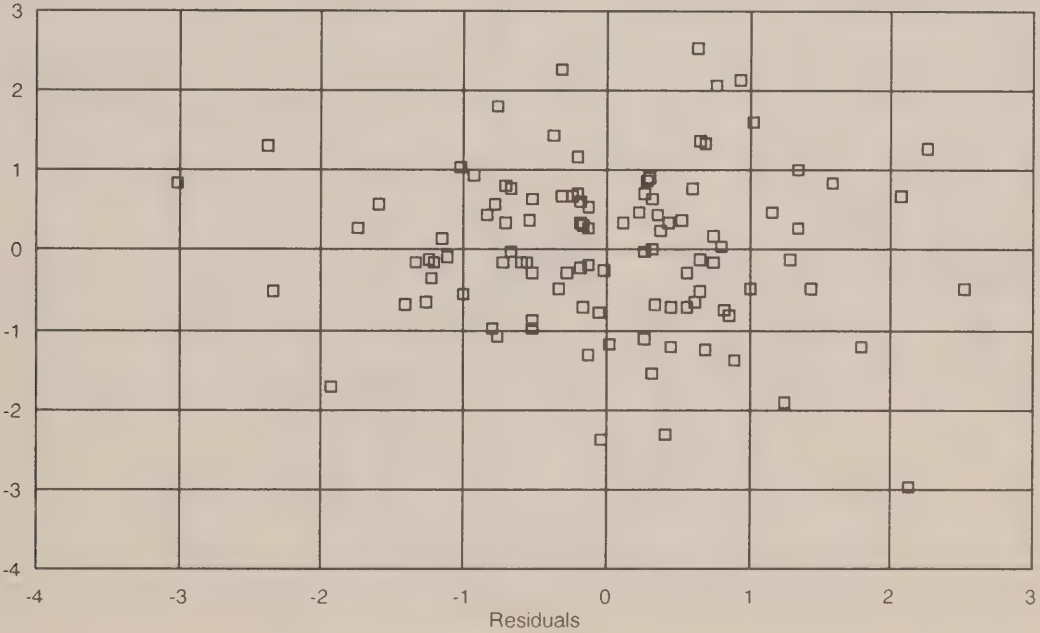
Graph 1b Cape Breton Island Observed and Smoothed Values (Log Transform)

Lagged residuals



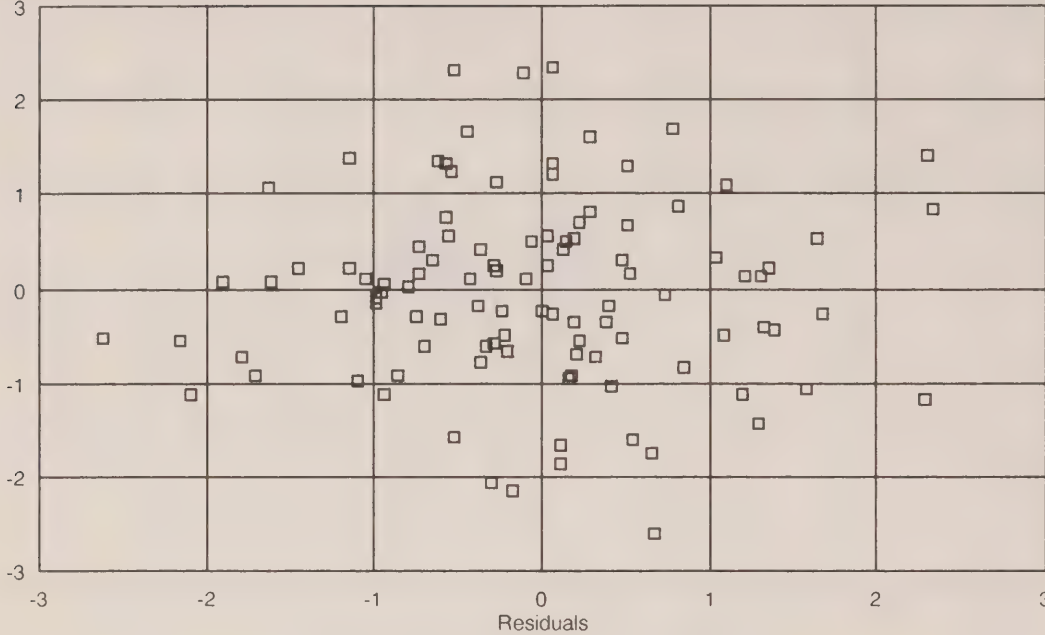
Graph 2a Nova Scotia One Step Ahead Prediction Errors – Survey Error Included

Lagged residuals



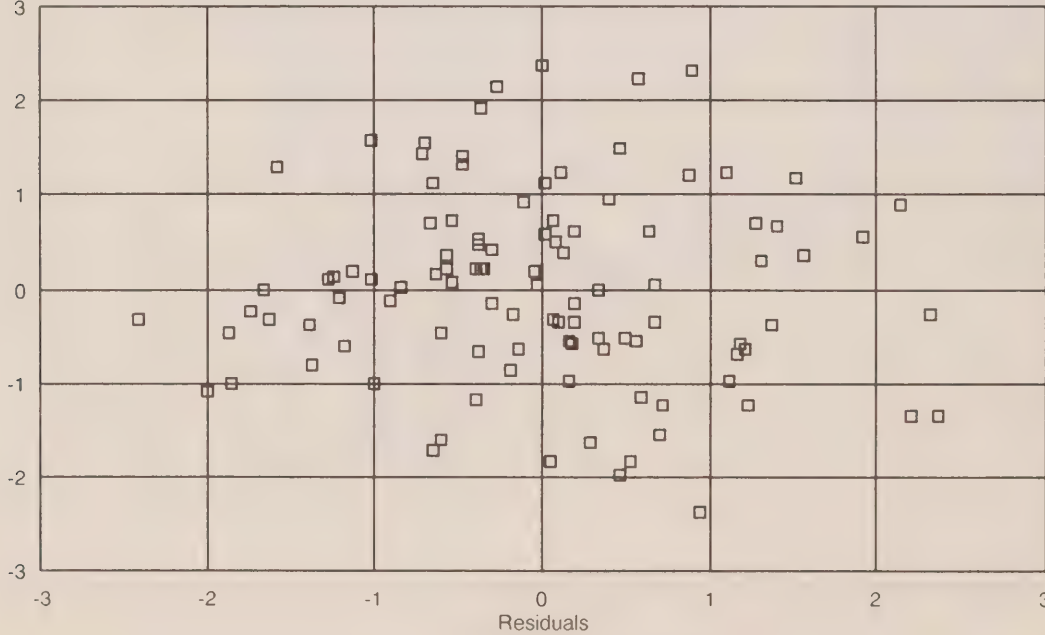
Graph 2b Nova Scotia One Step Ahead Prediction Errors – Survey Error Ignored

Lagged residuals

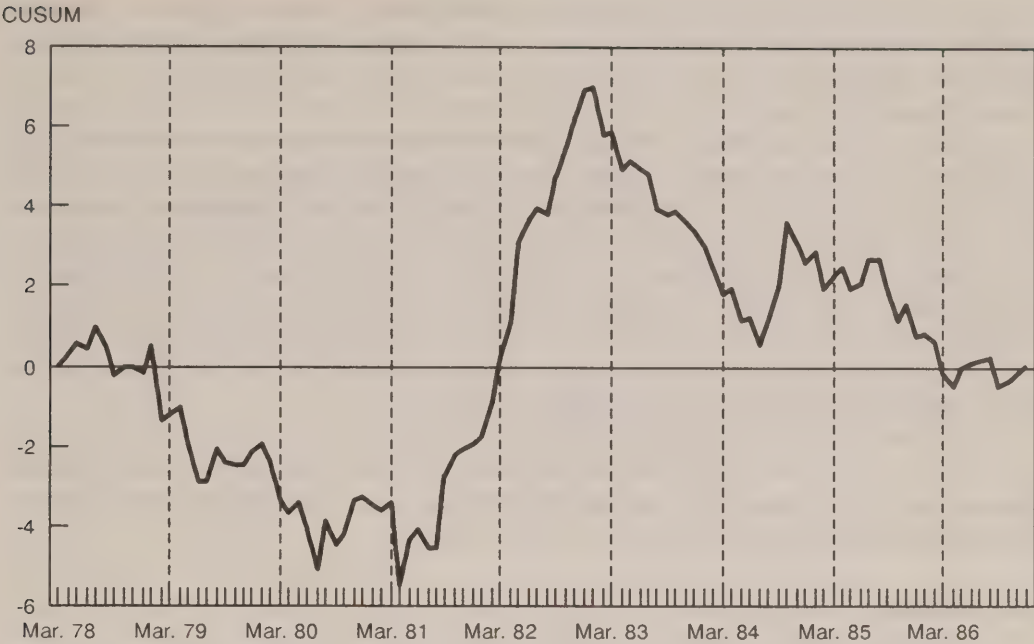


Graph 3a Cape Breton Island One Step Ahead Prediction Errors – Survey Error Included

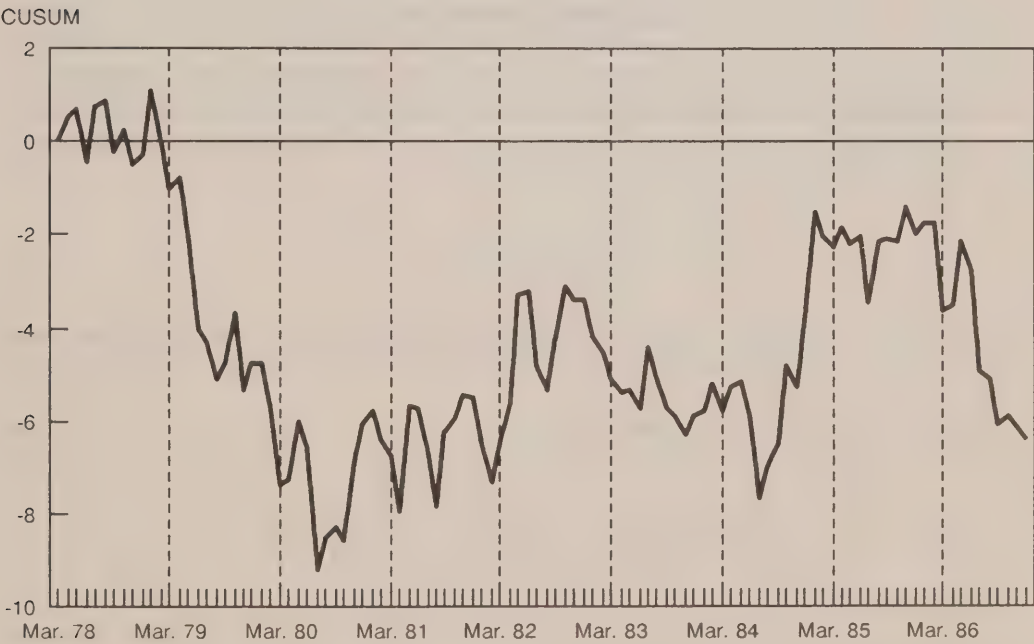
Lagged residuals



Graph 3b Cape Breton Island One Step Ahead Prediction Errors – Survey Error Ignored



Graph 4a Nova Scotia CUSUM of One Step Ahead Prediction Errors



Graph 4b Cape Breton Island CUSUM of One Step Ahead Prediction Errors

The same plots for Cape Breton Island are shown in Graph 3a and 3b. There is a striking similarity in the resulting residual plots for the two models from Cape Breton. However, none of the four plots give any compelling reason to doubt the underlying normal assumption of any of the models.

To test that the models did not undergo a structural change, the recursive residuals can be cumulatively summed to create a CUSUM chart. Whereas using the tests described in Brown, Durbin and Evans (1975) produced no significant results, the chart does suggest some structural change may be occurring. The CUSUM for Nova Scotia, as displayed in Graph 4a, shows quite clearly that prior to the recession the residuals are generally negative, implying that the model predictors are too large. During the 1981 recession the model produces mainly positive residuals. This implies that the model predictors are too small. The CUSUM for the Cape Breton Island models is shown in Graph 4b. Here we can see that the model that includes the survey error undergoes an earlier structural change.

We see, therefore, that model improvements can be made. By incorporating an extra regression variable corresponding to the structural changes noted in the CUSUM chart, further analysis can be performed within the same general framework. The form of such a variable is currently being investigated.

5.3 Summary

These examples demonstrate the importance of accounting for survey errors in certain time series analyses. Using the modified Kalman filter, we have developed a flexible method for parameter estimation, data smoothing and model validation for a wide variety of commonly used models.

ACKNOWLEDGEMENTS

The authors are grateful to Bill Steele of Social Survey Methods Division for his work in programming the algorithm described in Section 4. We are also grateful to the Labour Force Survey for supplying the data series. The referee and the associate editor both supplied many useful comments that improved the paper.

REFERENCES

- ANSLEY, C.F., and KOHN, R. (1985). A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. *Journal of Statistical Computation and Simulation*, 21, 135-169.
- ANSLEY, C.F., and KOHN, R. (1986). Prediction mean squared error for state space models with estimated parameters. *Biometrika*, 73, 467-473.
- BELL, W.R., and HILLMER, S.C. (1987). Time Series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.
- BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-68.
- BROWN, R.L., DURBIN, J., and EVANS, J.M. (1975). Techniques for testing the consistency of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37, 149-163.

- HARVEY, A.C., and DURBIN, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society, Series A*, 149, 187-222.
- HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- HAUSMAN, J.A., and WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.
- KOHN, R., and ANSLEY, C.F. (1986). Estimation, prediction and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, 81, 751-761.
- LEE, H. (1990). Estimation of panel correlations for the Canadian Labour Force Survey. *Survey Methodology*, 16, 283-292.
- MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. *Ph. D. Thesis*, Iowa State University, Ames, Iowa.
- RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kaltin and M.P. Singh), New York: Wiley, 457-479.
- SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F., and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys, *International Statistics Review*, 45, 13-28.
- TUNNICLIFFE-WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society, Series B*, 51, 15-27.

Spatial-Temporal Modelling of Spatially Aggregate Birth Data

DAVID R. BRILLINGER¹

ABSTRACT

Births by census division are studied via graphs and maps for the province of Saskatchewan for the years 1986-87. The goal of the work is to see how births are related to time and geography by obtaining contour maps that display the birth phenomenon in a smooth fashion. A principal difficulty arising is that the data are aggregate. A secondary goal is to examine the extent to which the Poisson-lognormal can replace for data that are counts, the normal regression model for continuous variates. To this end a hierarchy of models for count-valued random variates are fit to the birth data by maximum likelihood. These models include: the simple Poisson, the Poisson with year and weekday effects and the Poisson-lognormal with year and weekday effects. The use of the Poisson-lognormal is motivated by the idea that important covariates are unavailable to include in the fitting. As the discussion indicates, the work is preliminary.

KEY WORDS: Aggregate data; Borrowing strength; Contouring; Extra-Poisson variation; Locally-weighted analysis; Maps; Periodogram; Poisson distribution; Poisson-lognormal distribution; Random effects; Spatial data; Time series; Unmeasured covariates.

1. INTRODUCTION

The concern of this work is spatial-temporal data, that is quantities recorded as functions of space and time. The analysis of such data should be “easy” because of the graphing possibilities, *e.g.* rate versus time or effect versus geography, in the manner of residual plots so often employed in regression analysis; however in the present case the aggregation of basic elements leads to substantial difficulties.

The specific data studied consists of daily births for the calendar years 1986 and 1987 to women aged 25-29 for each of the 18 census divisions of the province of Saskatchewan. The corresponding population sizes, as determined in the 1986 Census, are also employed in order to compute rates. The reason that Saskatchewan was selected for this pilot study is that it is moderate sized and its boundaries and those of its census divisions are fairly regular. (The latter was important at the early stages of the work because computer based maps were then unavailable). Women aged 25-29 were selected because that was the 5 year age group with most births. These data were provided to the author by Statistics Canada. They are characterized by being aggregate, by being non Gaussian and by being non stationary in space and time.

It is wished to understand the relationship of births to time and geography, specifically to allow temporal and spatial patterns of fertility and possible surprises to show themselves. There are two central aspects to the study; a locally-weighted analysis of aggregate data is developed and random effects models are set down and fit to handle extra-Poisson variation. The latter part may be viewed as an inquiry into the flexibility of the Poisson-lognormal to handle unmeasured covariates and errors. The locally weighted analysis proceeds by developing weights, $w_i(x,y)$, that are meant to reflect the influence of the i -th census division (an aggregate) on the point location with coordinates (x,y) . Given census division data, these

¹ David R. Brillinger, Statistics Department, University of California, Berkeley, CA, 94720, U.S.A.

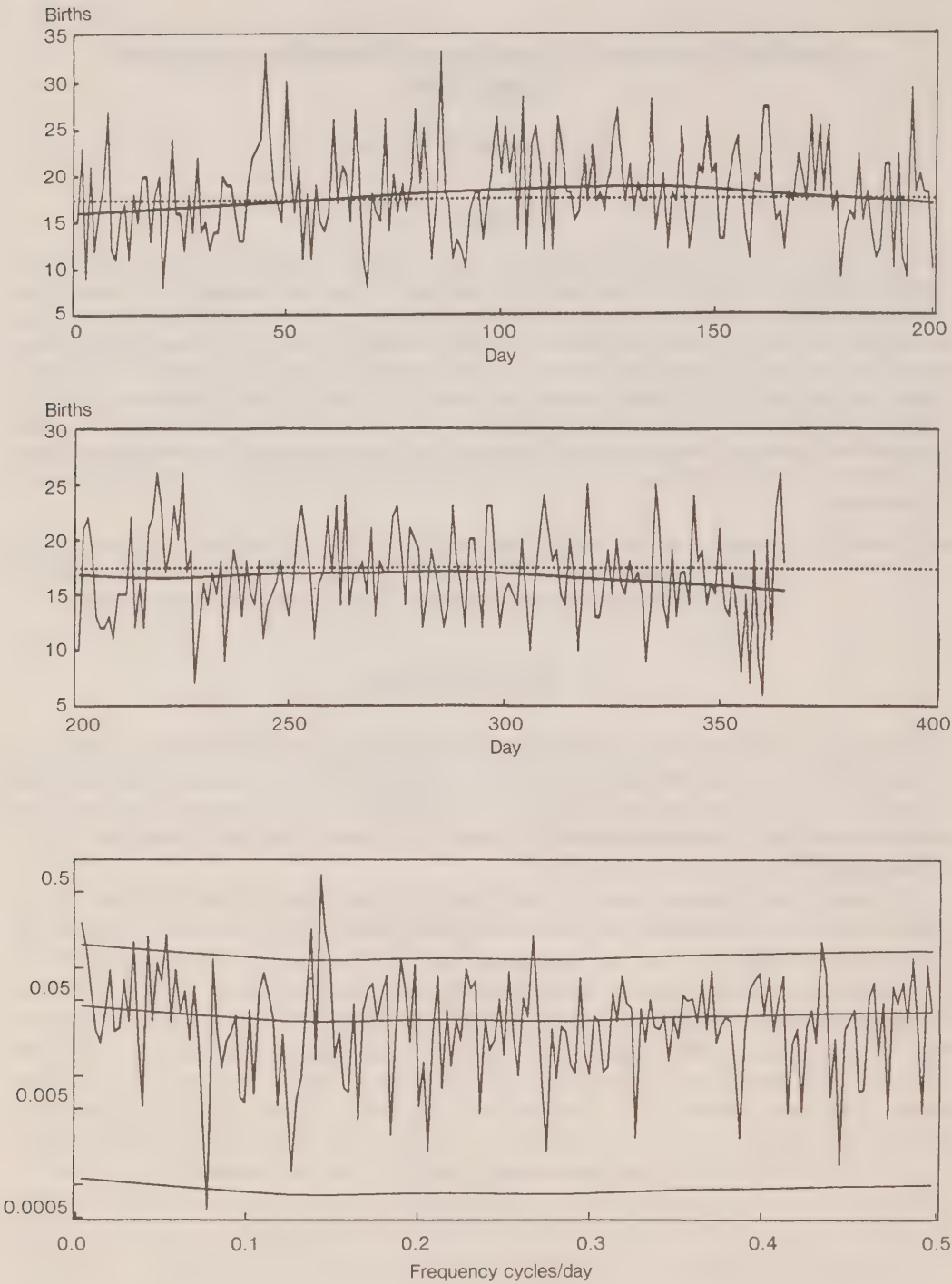


Figure 1. Top: Time series of annual births to women aged 25-29 in 1986 for the Province of Saskatchewan. Bottom: Periodogram of the square roots of the count graphed above. The solid lines provide approximate 95% marginal confidence limits. The peak corresponds to a period of 7 days.

weights are then applied to individual terms of the log-likelihood or corresponding estimation equations and parameter estimates evaluated.

It is to be emphasized that this is a preliminary report on work in progress. For example the fine structure of the data is not taken advantage of and no measures of uncertainty of the various estimates have been provided. The expressions employed for the weights, in this present work, are naive and bound to change form with further study, but the character of the analysis may be anticipated to remain of some interest.

The companion paper Brillinger (1990) considers some aspects of the spatial case alone.

2. BIRTHS AS A TIME SERIES

The top graph of Figure 1 provides the total number of births in Saskatchewan for each day of 1986. The dashed line is the 1986 mean level. The solid line is the result of heavily smoothing the series and is meant to highlight any trend. This graph does not, with casual inspection, provide striking evidence of any special phenomenon. However when the periodogram of the square root of the counts is computed, see bottom graph of Figure 1, something of interest appears. (The square root is employed to make the series more nearly symmetrical and normal). The upper and lower solid lines on the graph provide approximate 95% marginal confidence limits about a heavily smoothed version. A peak is apparent at a frequency of .143 cycles/day corresponding to a period of 7 days. This periodic phenomenon is well known in the analysis of birth data, see *e.g.* Cohen (1983) and Miyaoka (1989) and references therein. It is usually ascribed to doctors intervening in the natural process of labour and inducing births particularly on weekdays.

3. BIRTHS AS A SPATIAL PROCESS

Figure 2 provides, for each census division, and for women aged 25 to 29 the annual rate of births for the years 1986 and 1987 combined. One sees the highest rate of .208 births per woman per year to occur in the northern half of the province while the two lowest rates appear in the census divisions containing Regina and Saskatoon.

Figure 3 provides the numerical difference between the annual rate for 1987 and that for 1986 for each of the 18 census divisions. (Note that the 1986 census population has been taken as the divisor in each case). The differences are scattered around 0. It is to be noted that these rates are, however, based on fairly widely varying population sizes.

In the previous section the presence of a phenomenon of period 7 days was noted. Figure 4 presents the difference between the average weekday rate and the average weekend rate, (weekdays meaning Monday through Friday) for each census division. In all but one census division, the weekday rate is higher. This is consistent with various other studies and, as suggested in Section 2, is very possibly due to doctors inducing labor on weekdays (to avoid births on weekends).

The various rates presented in Figures 2, 3, 4 are average values for individual census divisions.

4. PROBLEMS ARISING

Maps of most quantities of direct interest that assign average values to the wholes of counties thereby lie, lie, lie.

With these graphic words Tukey (1979) deplores the use of maps such as those of Figures 2, 3, 4 that are constant across geographic divisions. Indeed examination of Figure 2, as does common knowledge, suggests that the birth phenomenon quite likely varies smoothly across census division boundaries. A principal concern of this work is to develop contour maps displaying smooth variation. It is hoped that such maps will prove useful in the discovery of general stochastic descriptions of the phenomenon and will allow insightful exploratory analyses.

A second concern of this work is with the statistical distribution of the counts themselves. A natural special stochastic model to employ is the Poisson. Yet in past studies the birth process has been found to relate to many socio-economic quantities, *e.g.* diet, lifestyle, weather, environment, weekday, holidays, age structure. Further the population of the various census divisions has varied around the Census Day values throughout 1986-87 and lastly the women's ages are scattered from 25 to 29. In summary it seems necessary to employ a more flexible model than the Poisson, specifically a model able to handle omitted covariates. The Poisson-lognormal will be employed in this work. As a sideline, due to the presence of the standard deviation parameter in the Poisson-lognormal, there will be a borrowing of strength that takes place in combining the data values, in the manner described by Mallows and Tukey (1982). (The term "borrowing strength" is employed, rather than for example "empirical Bayes" as some might prefer, because it has been in use for a substantial time period and because of its broader implications). Dean *et al.* (1989) is another recent reference concerned with handling extra-variation.

5. LOCALLY-WEIGHTED ANALYSIS

In the case of nonaggregate data, locally-weighted fitting is a convenient fashion by which to estimate smoothly varying quantities. Suppose one has a variate Y with probability distribution $p(Y | \Theta)$ depending on the finite dimensional parameter Θ . Suppose one wishes an estimate of Θ particular to the location with coordinates (x, y) . Suppose the datum Y_i is available for location (x_i, y_i) . Let $W_i(x, y)$ be a weight dependent on the distance of (x_i, y_i) to (x, y) .

Consider estimating Θ by maximizing the weighted log-likelihood

$$\sum_i W_i(x, y) \log p(Y_i | \Theta) \quad (1)$$

or (often equivalently) by solving the system of estimating equations

$$\sum_i W_i(x, y) \Psi(Y_i | \hat{\Theta}) = 0 \quad (2)$$

with $\Psi(Y | \Theta) = \partial \log p / \partial \Theta$, the score function.

To illustrate the technique consider an elementary case, specifically take Y to be normal with mean μ and variance σ^2 . The locally weighted estimate of μ at (x, y) results from minimizing

$$\sum_i W_i(x, y) [Y_i - \mu]^2$$

and is given by

$$\hat{\mu}(x, y) = \sum_i W_i(x, y) Y_i / \sum_i W_i(x, y),$$

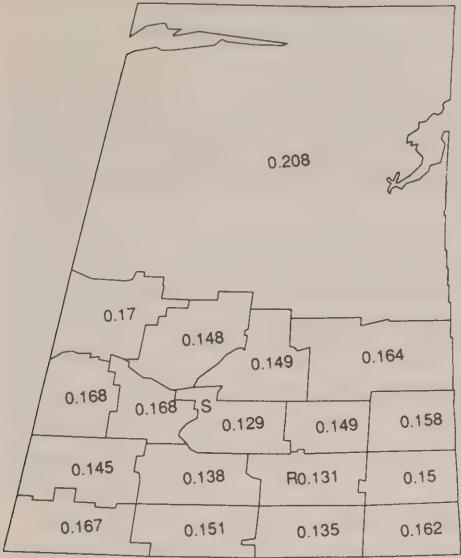


Figure 2. The average annual birth rate for women aged 25 to 29 for the years 1986 and 1987, plotted above census divisions. “R” and “S” indicate the locations of Regina and Saskatoon respectively.

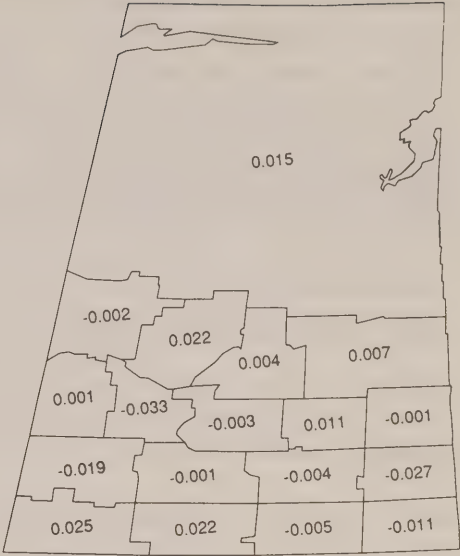


Figure 3. The 1987 rate minus the 1986 rate for the same data as Figure 2.

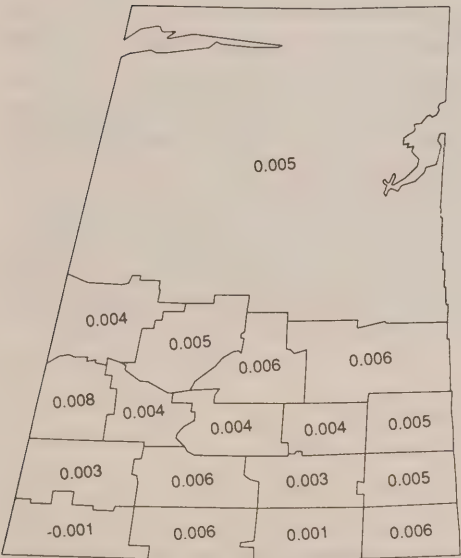


Figure 4. The average weekday rate minus the average weekend rate for the same data as Figure 2.

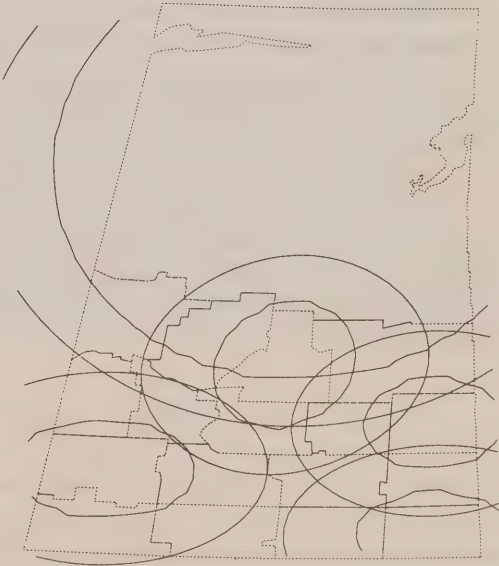


Figure 5. The weights, $W_i(x,y)$ applied in equations (1) or (2), computed via expression (4), for four of the census divisions. The weights are not shown for all the divisions in the interests of clarity. The contours at levels .50 and .99 are shown.

an expression with intuitive appeal. It is to be noted that such formulas are commonly used in computer graphics as interpolation procedures, see for example Franke (1982).

Among references we may mention Gilchrist (1967) concerned with "discounting", Pelto *et al.* (1968), concerned with least squares, Cleveland and Kleiner (1975), who suggested the use of moving midmeans and Stone (1977) focusing on regression. In the discussion of Stone's paper, Brillinger (1977) suggested the form (2) for a general distribution and justified it as a Bayes' rule. Specifically consider the loss function

$$L(Y | \Theta) = -\log p(Y | \Theta).$$

Suppose an estimate is desired at $\mathbf{r} = (x, y)$. The Bayes' risk may be written

$$E\{L(Y | \Theta_{\mathbf{r}})\} = E\{E\{L(Y | \Theta_{\mathbf{r}}) | \mathbf{r}\}\}.$$

Bayes' rule seeks

$$\min_{\Theta} E\{L(Y | \Theta) | \mathbf{r}\}.$$

With data Y_i, \mathbf{r}_i , and $W_i(\mathbf{r})$ a kernel centred at \mathbf{r}_i , one approximates the conditional expected value here by

$$E\{\log p(Y | \Theta) | \mathbf{r}\} \approx \sum_i W_i(\mathbf{r}) \log p(Y_i | \Theta)$$

and so is led to expression (1).

Tibshirani and Hastie (1987) develop an equi-weighted local likelihood estimation procedure. Cleveland and Devlin (1988) develop the least squares approach in real detail. Staniswalis (1989) studies and implements the general p case. Advantages of the locally-weighted technique include: no "hidden model" distribution assumption, the possibility of discerning non-additivity, variants for resistance and influence, simple additivity of the observation component, and no matrix inversion (as, for example, kriging requires).

The birth data of concern in this work is aggregate (or grouped) totals over census divisions. The procedure of the preceding section cannot therefore be employed directly. The problem is that of obtaining appropriate weights $w_i(x, y)$ evidencing the effect of the census division i on the location (x, y) . Suppose $|R_i|$ denotes the area of census division i . Then the naive weight function is

$$w_i(x, y) = \frac{1}{|R_i|} \quad \text{for } (x, y) \text{ in } R_i$$

and equal 0 otherwise. In this work functions of the essential form

$$w_i(x, y) = \frac{1}{|R_i|} \int_{R_i} W(x - u, y - v) du dv \quad (3)$$

will be employed where $W(\cdot)$ is a kernel appropriate for the nonaggregate case as for example studied in Cleveland and Devlin (1988). The formula (3) may be motivated by consideration of the Poisson point process case, see Appendix II. Estimates will be determined via the criteria (1) or (2) with W_i replaced by w_i .

The specific weights employed at $\mathbf{r} = (x,y)$ in this preliminary work are

$$w_i(\mathbf{r}) = \exp\{ - (1 - \rho)^2 \|\mathbf{r} - \mathbf{r}_i\|^2 / 2\tau^2 \} \tag{4}$$

outside the ellipse $(\mathbf{r}_0 - \mathbf{\bar{r}}_i) \mathbf{S}_i^{-1} (\mathbf{r}_0 - \mathbf{r}_i)' = d_0^2 = 5.991$ and equal 1 inside. Here $\|\mathbf{r}\|^2 = x^2 + y^2$, $\rho = d_0 / \sqrt{(\mathbf{r} - \mathbf{\bar{r}}_i) \mathbf{S}_i^{-1} (\mathbf{r} - \mathbf{\bar{r}}_i)'}$ and $\tau = .025$, where $\mathbf{\bar{r}}_i = E U_i$ and $\mathbf{S}_i = \text{var } U_i$ with U_i a variate uniformly distributed within R_i . This choice of ρ makes the weight function continuous. The logic is that the census divisions are approximated by ellipses with the same mean and variance-covariance matrix. (The specific values were chosen after a bit of experimentation, in part to make the area in the initial ellipse about .95 of the division's). One could have employed other shapes than ellipses, *e.g.* rectangles, but this is preliminary work and it is anticipated that later work will employ weights of the form (3).

Figure 5 displays the .50 and .99 contours of the $w_i(x,y)$ plotted for several of the census divisions. The contours are seen to follow the general shapes of the census divisions. The jaggedness in some of the contours results from the discreteness of the 40×40 grid employed in the computations.

Other weight functions constructed with somewhat similar problems in mind may be found in Tobler (1979) and Dyn and Wahba (1982). Advantages of the present approach, as listed for the nonaggregate case above include: the terms in (1) or (2) are additive and do not interact, no matrix inversion is needed, and resistance to outliers is easily built in.

Cliff and Ord (1975) Section 5.1, discusses measures of the influence of counties on other counties. The concern of this present paper however is the influence of a "county" on a point location. It is to be remarked that perhaps the weight, providing the influence, should depend on some covariates, *e.g.* county population.

6. A POISSON FIT

Throughout the analysis, the female population aged 25-29 and births to its members will be considered. Let $i = 1, \dots, 18$ index census division. Let N_i denote the census count of the women in the i -th division. (These are the counts for Census Day, 3 June 1986). Let B_i denote the total number of births to women aged 25-29 in the two years 1986-87.

Suppose that the probability distribution $p(\cdot)$ of Section 5 is that B_i is Poisson with mean $2N_i\mu$. (The presence of the multiplier 2 is so the parameter μ is an annual birth rate). One logic for the Poisson assumption comes from the idea that birthdays are random, see Brillinger (1986).

With the Poisson assumption, the locally weighted estimate of the annual birth rate at location (x,y) is given by

$$\hat{\mu}(x,y) = \sum_i w_i(x,y) B_i / 2 \sum_i w_i(x,y) N_i. \tag{5}$$

These values are computed for (x,y) on a 40 by 40 grid and the corresponding contour plot is given in Figure 6. The contours are seen to vary smoothly. This (smoothed) rate varies from .14 to .20, with the higher values in the upper half and the lower centred around the Province's most urban part.

As indicated previously, the data under study has important temporal characteristics. Models need to take this into account. In particular the weekly periodicity needs to be handled as well as possible trends in population sizes. The following model seems worth considering. Let j be an indicator variable with $j = 1$ if the count is for a weekday and $j = 2$ if the count is for

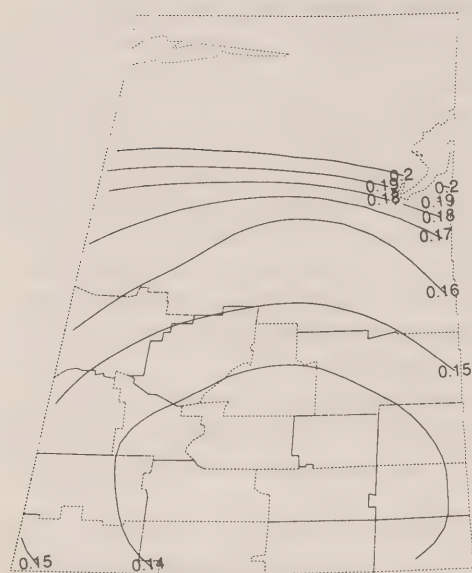


Figure 6. Expression (5) graphed for the weights of (4) with B_i the count of births in census division i during 1986-87 and N_i the corresponding population count of women aged 25-29.

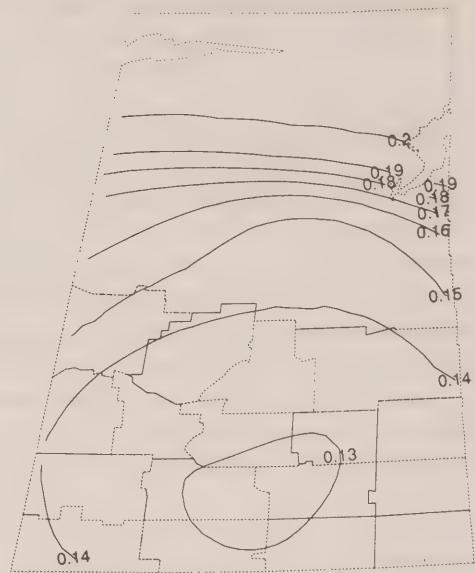


Figure 7. The estimated birth rate $\exp\{\hat{\alpha}\}$ obtained by locally weighted fitting assuming that the number of births, B , given the population at risk, N , is Poisson with mean $N\exp\{\alpha \pm \beta \pm \gamma\}$ with the first \pm sign plus for weekdays and minus for weekends and the second \pm plus for 1986 and minus for 1987.

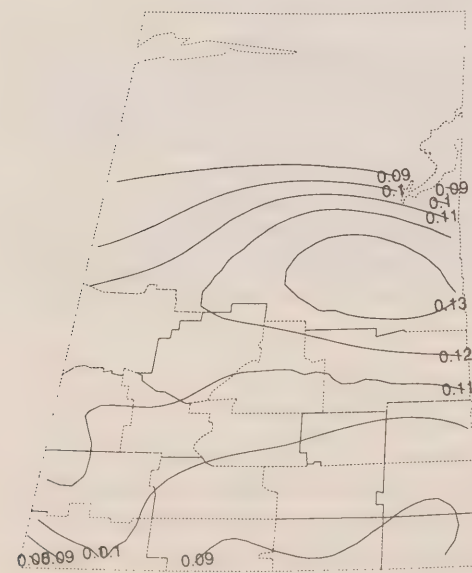


Figure 8. Plot of the estimated weekday effect $\hat{\beta}(x,y)$ obtained as per Figure 7.

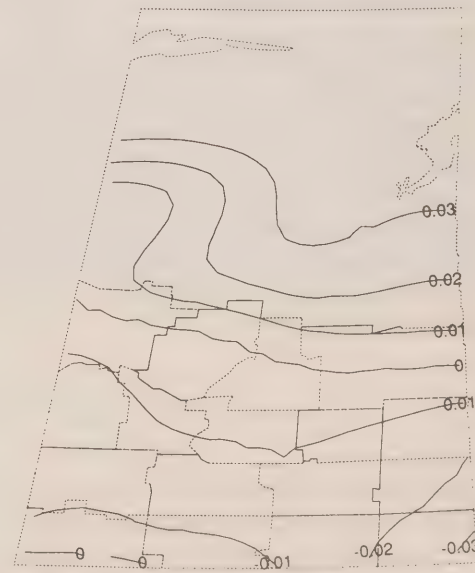


Figure 9. The estimated year effect $\hat{\gamma}(x,y)$ as per Figure 7.

a weekend. Let k be a second indicator variable with $k = 1$ for 1986 and $k = 2$ for 1987. Let B_{ijk} denote the corresponding number of births in census division i . Suppose that B_{ijk} given N_i is Poisson with mean $N_i \exp\{\alpha + \beta_j + \gamma_k\}$. β_j is the weekday effect, γ_k the year effect and it will be assumed that $\beta_1 + \beta_2, \gamma_1 + \gamma_2 = 0$ to make the model identifiable. If there is no weekday effect, then $\beta_1, \beta_2 = 0$. If there is no year effect, then $\gamma_1, \gamma_2 = 0$. Now, via locally-weighted analysis presented in Section 5, one can obtain estimates of α, β and γ as functions of location (x, y) . (For simple balance in the computations, only the first $364 = 7 \times 52$ days of each year have been employed).

Figure 7 provides the estimate $\exp\{\hat{\alpha}(x, y)\}$ obtained of the annual birth rate. It is interesting to note that, relative to the constant rate Poisson model, the contours have expanded out somewhat from the urban areas. Figure 8 provides the estimated weekday effect, $\hat{\beta}_1(x, y)$, obtained. In its case there is bulge to the east. These values are quite a different representation from that of the naive differences of Figure 4. In particular, now there is a reflection of the differing population sizes. The order of magnitude of the $\hat{\beta}$'s is .08 to .13 while $\hat{\alpha}$ is order -2.1 to -1.6 . Figure 9 provides the estimated year effect, $\hat{\gamma}_1(x, y)$. Its values vary from $-.03$ to $.03$. Numerically, the weekday-weekend effect is the larger.

The just preceding analysis suggests that there are basic variables that can affect birth rates and that modelling and analysis needs to take this circumstance into account.

7. POISSON-LOGNORMAL FITS

With a multi-dimensional explanatory variable x in hand, a Poisson model that has B of mean $N \exp\{x\theta\}$ might do a good job of explaining the data. Examples of explanatory variables include: diet, lifestyle, weather, environment, holidays, population change, age structure, vagaries of boundaries. In the present situation, these variables are not at hand. The omitted variables in the model will be assumed specifically accumulated into an error variable. It will be assumed that, given ϵ , the variate B is Poisson with mean $N\mu \exp\{\epsilon\}$ and that ϵ is normal with mean 0 and variance σ^2 . In the case of this model B is said to have a Poisson-lognormal distribution. Some information on this distribution may be found in Shaban (1988). Sometimes ϵ enters directly from the problem context, see Brillinger and Preisler (1983) for one example, but in the present case it is simply assumed present.

A critical difficulty, that arises in working with a Poisson-lognormal model, is that closed expressions do not exist for the probability function. Yet the model is clearly flexible for introducing effects and handling unavailable variables. Following the work of Bock and Lieberman (1970), Pierce and Sands (1975) and Hinde (1982), one can proceed via numerical quadrature. The probability function may be written

$$p(Y) = \frac{1}{Y!} \int (ve^{\sigma z})^Y \exp\{-ve^{\sigma z}\} \phi(z) dz$$

with ϕ the standard normal density, with Y corresponding to B and with v corresponding to $N\mu$. To proceed with a data analysis the integral is approximated by a finite number of terms involving nodes, z_l , and weights, w_l ,

$$p(Y) \approx \frac{1}{Y!} \sum_{l=1}^L (ve^{\sigma z_l})^Y \exp\{-ve^{\sigma z_l}\} w_l.$$

Listings of nodes and weights may be found in Abramowitz and Stegun (1964) for example.

Figures 10, 11, 12, 13 provide the results of fitting the Poisson-lognormal model including weekday and year effects and employing $L = 5$ nodes. The model assumes B_{ijk} given N_i and Z is Poisson with mean

$$N_i \exp\{\alpha + \beta_j + \gamma_k + \sigma Z\}$$

Z denoting a standard normal deviate and further assumes the separate Z 's independent. Here i indexes census division, j weekday or not and k year. Figure 10, a contour plot of $\exp\{\hat{\alpha}(x,y)\}$, again shows a dip around the urban region as in Figure 7. The irregularity in the figure suggests that in one case perhaps the estimation procedure converged to a local extremum. Figures 11 and 12 similarly provide $\hat{\beta}(x,y)$ and $\hat{\gamma}(x,y)$. There are again suggestions of local extrema. Figure 13, a contour plot of $\hat{\sigma}(x,y)$, is not easily described. It suggests that the estimate, $\hat{\sigma}$, is fairly variable. The estimate is seen to be of order of magnitude .1 and so comparable to the weekday effect of Section 6.

All the work on estimation with the Poisson-lognormal, that we know about, involves some form of approximation. For example Clayton and Kaldor (1987) approximate the conditional Poisson log-likelihood by a quadratic and Aitchison and Ho (1989) also employ numerical integration, albeit after a transformation of the parameters. A new type of approximation has recently been proposed in Crouch and Spiegelman (1990). Its effectiveness for the Poisson-lognormal remains to be studied.

8. DISCUSSION

Locally-weighted analysis and random effect models appear to provide a flexible means of dealing with a broad class of problems involving geographic data. The random effect terms have two important roles: handling omitted effects and borrowing strength for improved estimates of the principal parameters. For the Poisson alone, naive totals are efficient, yet there exists extra-Poisson variability due to omitted variables in the present case.

The approach is computer intensive, because of the numerical integration and the maximum likelihood estimation at many points on a grid, but proved quite manageable on the Berkeley network of Sun 3/50's.

Much future work remains including: tools for assessing fit, uncertainty computation and display, weight function choice (particularly choice of τ in (4)), analyses for other age groups and provinces, and appropriate asymptotics. Further understanding needs to be gained as to why with nearby initial values the optimizing routine sometimes converged to somewhat distant estimates. An advantage of the present circumstance is that there exists immense amounts of other data to be made use of as work progresses. Examination of Figures 6 on shows an important limitation of the technique – it is providing too much fine detail in the northern half of the province.

Other recent papers devoted to the analysis of vital statistics rates are: Cressie and Read (1989), Clayton and Kaldor (1987), Tsutakawa (1988) and Manton *et al.* (1989). These papers are however not directed at the problem of obtaining a smooth surface, which is the concern of this work.

It is amusing to note that the presence of the weekly period in the phenomenon allowed the author to deduce early on in the work that a confusion had arisen over which data set was to be supplied. When the days of fewest births were determined for the initial data set supplied, the days were found to be (apparently) Friday and Saturday. This was because the year 1987 had been supplied, and not the desired 1986.

After the analyses were completed it was learned that the birth counts were based on 1981 census divisions, while the population counts were based on 1986. Luckily the boundaries have not changed much, but this circumstance provides yet more reason for wanting a procedure that can handle extra-variation.

9. ADDENDUM

In the paper a case has been made for the inclusion of an error term, ϵ , to reflect pertinent covariates that were unavailable for the analysis. This led to the employment of the Poisson-lognormal distribution. In Tukey (1990) an index of urbanicity of a census division is constructed. It is based on the populations of the three largest places in the division. The values, x_i , of the index are given in Figure 14 and are seen to be lowest in the census divisions containing Regina and Saskatoon.

The table below gives the results of employing Glim to fit the successive Poisson models for B_{ijk} given N_i : (i) $N_i \exp\{\alpha + \beta_j + \gamma_k\}$, (ii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta x_i\}$, and (iii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta_1 x_i + \delta_2 x_i^2\}$.

Variables	Deviance	d.f.	<i>p</i> -value
weekday, year	227.3	69	
+ urbanicity	86.69	68	
+ urbanicity**2	83.13	67	.088

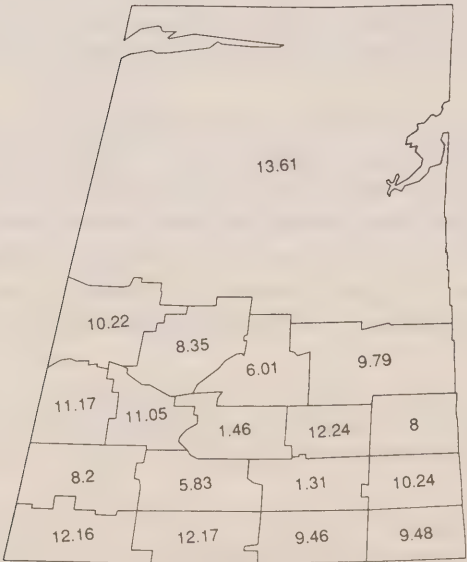


Figure 14. The values of the Tukey index of urbanicity.

By bringing in this urbanicity variable, x_i , now a Poisson model is satisfactory for the circumstance.

Finally the Referee made some comments that spell out quite specifically the assumptions and limitations of this present study. The work is continuing and the intention is to address these comments. Rather than paraphrasing, it seems more sensible to provide the referee's own words.

"The choice of weights is *ad hoc* and requires more thought. If one had two divisions, both of the same area but with vastly different populations N_i , should the weighting be the same? It depends on whether area or population density is thought to be more important. Use of the latter may remove the spurious fine detail in the northern half of the province."

"There are traps with N_i 's, which the author appears to be aware of, but I think the reader needs extra warning. It might help to have approximate measures of uncertainty ([Section 1] promises none). Figure 3 cannot really be interpreted, since positive or negative values may be due to random fluctuations about zero. The contours in Figure 6 are calculated with vastly different precision, and in some respects are incomparable. And, [in Section 6], upon estimating α , β and γ , it would be tempting (but unwise) to assume that such values are significant."

"All random variables in sight are assumed independent. Another way to motivate these weighted models is to assume a multivariate distribution, with the property that the conditional mean at (x,y) , given the surrounding data, is a weighted combination of those data. Then the joint distribution exhibits dependence."

ACKNOWLEDGEMENTS

The author would like to thank G. Brackstone, R. Gussella, R. Raby, B. Sander, P. Spector, M. Subhani, R. Villani for assistance in obtaining the data and maps, for help with computational geometry and with parallel computing. John Tukey, Rob Tibshirani and the Referee made very helpful comments on the first draft. The research was supported by National Science Foundation Grant DMS-8900613.

APPENDIX I

In this Appendix a few computing details are provided. The census divisions and the province boundaries are specified as polygons. To compute the weights $w_i(x,y)$ an algorithm was required to check whether a given point was inside a given polygon. To compute the mean and variance of a random point inside a given polygon, an algorithm for breaking a polygon up into triangles was required. Such algorithms are discussed in Preparata and Shamos (1985) for example. The approximate likelihood was maximized via the Harwell FORTRAN routine *va09a*. For the parallel computations the 40 by 40 grid was broken up into 20 disjoint segments and the computations thence carried out on 20 separate work stations. As in Brillinger and Preisler (1983), factors were introduced into the likelihood to stabilize the computations. Miyaoka (1989) found that the computations could be sensitive to the number of nodes employed. In the present series of computations, the number was increased until the results did not change much. There is also the problem of selecting initial values. Here they were taken to be the method of moment estimates, although these are perhaps too inefficient.

APPENDIX II

For simplicity, consider the case of a point process $\{x_j\}$ with rate function v on the line. The local weighted log likelihood for a Poisson process is, up to a constant,

$$\sum_j W(x - x_j) \log v(x_j) - \int W(x - u) v(u) du.$$

So, the locally weighted estimate of the rate is

$$\hat{v}(x) = \sum_j W(x - x_j) \bigg/ \int W(x - u) du,$$

the usual form of estimate. Suppose now the line is broken into intervals R_i , and the aggregate count available is $N(R_i)$. One desires

$$\sum_{x_j \in R_i} W(x - x_j).$$

If this last is to be approximated by $\Theta N(R_i)$, then the method of moments leads to

$$\Theta = \int_{R_i} W(x - u) du \bigg/ |R_i|$$

and thence to expression (3).

REFERENCES

- ABRAMOWITZ, M., and STEGUN, I.A. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington.
- AITCHISON, J., and HO, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.
- BOCK, R.D., and LIEBERMAN, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- BRILLINGER, D.R. (1977). Discussion of Stone (1977). *The Annals of Statistics*, 5, 622-623.
- BRILLINGER, D.R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42, 693-734.
- BRILLINGER, D.R. (1990). Mapping aggregate birth data. Technical report, Statistics Department, University of California, Berkeley.
- BRILLINGER, D.R., and PREISLER, H.K. (1983). Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*, 31-65. New York: Academic Press.
- CLAYTON, D., and KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.
- CLEVELAND, W.S., and DEVLIN, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.
- CLEVELAND, W.S., and KLEINER, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics*, 17, 447-454.

- CLIFF, A.D., and ORD, J.K. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society*, 37, 297-348.
- COHEN, A. (1983). Seasonal daily effect on the number of births in Israel. *Applied Statistics*, 32, 228-235.
- CRESSIE, N., and READ, T.R.C. (1989). Spatial data analysis of regional counts. *Biometrical Journal*, 31, 699-719.
- CROUCH, E.A.C., and SPIEGELMAN, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp\{-t^2\} dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.
- DEAN, C., LAWLESS, J.F., and WILLMOT, G.E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics*, 17, 171-182.
- DYN, N., and WAHBA, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal on Mathematical Analysis*, 13, 134-152.
- FRANKE, R. (1982). Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38, 181-200.
- GILCHRIST, W.G. (1967). Methods of estimation involving discounting. *Journal of the Royal Statistical Society*, 29, 355-369.
- HINDE, J. (1982). Compound regression models. GLIM 82 (Ed. R. Gilchrist), 109-121. *Lecture Notes in Statistics*, 14. New York: Springer-Verlag.
- MALLOWS, C., and TUKEY, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (Eds. Tiago de Oliveira, J. et al.), 111-172. London: Academic.
- MANTON, K.G., WOODBURY, M.A., STALLARD, E., RIGGAN, W.B., CREASON, J.P., and PELOM, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates. *Journal of the American Statistical Association*, 84, 637-650.
- MIYAOKA, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. *Canadian Journal Statistics*, 17, 123-140.
- PELTO, C.R., ELKINS, T.A., and BOYD, H.A. (1968). Automatic contouring of irregularly spaced data. *Geophysics*, 33, 424-430.
- PIERCE, D.A., and SANDS, B.R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Department, Oregon State University.
- PREPARATA, F.P., and SHAMOS, I. (1985). *Computational Geometry*. New York: Springer-Verlag.
- SHABAN, S.A. (1988). Poisson-lognormal distributions. *Lognormal Distributions*, 195-210 (Eds. E.L. Crow and K. Shimizu). New York: Marcel Dekker.
- STANISWALIS, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.
- STONE, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595-620.
- TIBSHIRANI, R., and HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.
- TOBLER, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519-536.
- TSUTAKAWA, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83, 37-42.
- TUKEY, J.W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Included in The Collected Works of J.W. Tukey, Vol. 5 (1988), (Ed. W.S. Cleveland). Pacific Grove: Wadsworth.
- TUKEY, J.W. (1990). Graphical displays of: Are the (x,y) pairs compatible with a linear dependence? Technical Report No. 301, Princeton University.

Benchmarking of Economic Time Series

NORMAND LANIEL and KIMBERLEY FYFE¹

ABSTRACT

Benchmarking is a method of improving estimates from a sub-annual survey with the help of corresponding estimates from an annual survey. For example, estimates of monthly retail sales might be improved using estimates from the annual survey. This article deals, first with the problem posed by the benchmarking of time series produced by economic surveys, and then reviews the most relevant methods for solving this problem. Next, two new statistical methods are proposed, based on a non-linear model for sub-annual data. The benchmarked estimates are then obtained by applying weighted least squares.

KEY WORDS: Survey errors; Non-linear model; Weighted least squares.

1. INTRODUCTION

Traditionally benchmarking has been defined as the method of adjusting monthly or quarterly figures derived from one source to annual values (benchmarks) obtained via another source (see Denton 1971, Cholette 1988a, and Monsour and Trager 1979). For example, the monthly shipments of Canadian Manufacturers could be adjusted so that they add up to the Annual Census of Manufacturers shipments figures. Another definition of benchmarking is the more general one of improving sub-annual estimates derived from one source with annual estimates obtained via a second source (see Hillmer and Trabelsi 1987). This definition assumes that the annual values are subject to error, which is not the case with the first definition. For example, the monthly inventories of Canadian Retailers derived from a sample survey could be improved using the end of year inventories obtained from the annual retail trade sample survey. This second definition of benchmarking corresponds to the situation encountered with many economic time series and is the one dealt with in this paper.

The purpose of this article is twofold, first it describes in detail, the benchmarking problem as it appears for many time series produced by large scale economic surveys. Then, two well known benchmarking methods dealing with a single time series are presented and discussed. Since both of these methods fail in some respects to resolve the problem, two other methods which use a non-linear weighted least squares approach are proposed. Finally, two of the above mentioned methods are illustrated with some simulated data and the results are discussed.

2. PROBLEM DESCRIPTION

The problem of improving a two-way table of sub-annual series of estimates with annual series of estimates from business surveys is described here, accompanied by the characteristics of the original data and a list of the features desired from a benchmarking procedure.

¹ Normand Laniel and Kimberley Fyfe, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The sub-annual estimates are often biased due to frame coverage deficiencies. Undercoverage is caused by delay in the inclusion of new businesses and no representation of non-employer businesses (usually small ones) on the frame. These sub-annual estimates are usually derived from relatively small overlapping samples, implying that sampling variances are relatively large and that sampling covariances exist between sub-annual estimates of different time periods. In addition, most economic sub-annual surveys produce series of estimates for a number of industrial activities within a number of geographical regions. These are published sub-annually in the form of industry by geographical region tables, where the cells as well as the marginals and the grand totals need to be benchmarked.

As regards annual estimates, they can be assumed to be unbiased since in practice their frames do not suffer much from coverage deficiencies. Also, the annual estimates usually come from relatively large samples or censuses and thus have relatively small or no sampling errors associated with them, while their sampling covariances tend to be large because of substantial sample overlap between years. Another point to note about the annual estimates is that these figures come in approximately two years after the time to which they refer. For example, annual data for 1988 will not be released until some time in 1990, while sub-annual data are usually available a few months after the time period to which they refer. Therefore, when the sub-annual estimates are to be benchmarked, there will be no annual benchmarks for some of the sub-annual periods.

There are a number of features that a benchmarking procedure should have in order to be used for large scale survey estimates. First, the procedure should be simple enough that it can be used without too much data analysis. Second, it must be possible to produce preliminary benchmarking factors for periods for which benchmarks are not yet available. This feature allows benchmarking to be performed as the sub-annual data are produced. Otherwise discontinuities will be introduced in the sub-annual data. It is also desirable that the method maintain consistency between the grand-totals, marginal totals, and cell estimates for the benchmarked estimates in a table.

More discussion on the last two features can be found in Laniel and Fyfe (1989) and (1990) and Cholette (1988a) and (1988b). The rest of this paper deals with the problem of benchmarking a single time series in the context described above.

3. BENCHMARKING A SINGLE SERIES

Four approaches to benchmarking a single time series of sub-annual flow or stock estimates are described in the following sub-sections.

3.1 Denton's Method

In his 1971 paper, Denton proposed procedures for benchmarking based on a Quadratic Minimization approach, each of which corresponds to a specific penalty function. One of these penalty functions is the proportionate first difference between the original and benchmarked series and is often used for the problem of benchmarking time series that was described in section 2. Denton's procedure can be presented in statistical terms by first stating that the sub-annual estimates follow the model:

$$\frac{\theta_t}{y_t} = \frac{\theta_{t-1}}{y_{t-1}} + \epsilon_t, \quad t=1, 2, \dots, n \quad (3.1)$$

subject to the restriction to the annual data:

$$z_T = \sum_{t \in T} \theta_t, \quad T = 1, 2, \dots, m, \quad (3.2)$$

where:

- t refers to a sub-annual period,
- T refers to an annual period,
- $\{y_t\}$ is a sequence of biased estimates of the sub-annual parameters (levels),
- $\{\theta_t\}$ is a sequence of fixed sub-annual parameters (true values of the levels),
- $\{\epsilon_t\}$ is a sequence of uncorrelated and identically distributed errors with mean vector and covariance matrix $(\mathbf{0}, \sigma^2 I)$ and,
- $\{z_T\}$ is a sequence of annual benchmarks.

To find the benchmarked estimates, least squares are applied to the above restricted model.

It is important to note that Denton's approach assumes that the bias follows a random walk and that both the sub-annual and annual data are observed without sampling errors. Unfortunately, these assumptions are unlikely to be satisfied by economic time series (see section 2).

3.2 Hillmer and Trabelsi's Method

In 1987, Hillmer and Trabelsi proposed an alternative approach to benchmarking based on the Box-Jenkins (1976) ARIMA models. They assumed that the sub-annual estimates follow the model:

$$y_t = \theta_t + u_t \quad t = 1, 2, \dots, n \quad (3.3)$$

and the annual estimates follow the model:

$$z_T = \sum_{t \in T} \theta_t + a_T \quad T = 1, 2, \dots, m, \quad (3.4)$$

where:

- $\{\theta_t\}$ is a sequence of stochastic sub-annual parameters (true values of levels) following an ARIMA model,
- $\{y_t\}$ is a sequence of unbiased estimates of the sub-annual parameters,
- $\{u_t\}$ is a sequence of sub-annual dependent sampling errors with mean vector and covariance matrix $(\mathbf{0}, \Sigma_u)$,
- $\{z_T\}$ is a sequence of annual unbiased estimates, and
- $\{a_T\}$ is a sequence of annual dependent sampling errors with mean vector and covariance matrix $(\mathbf{0}, \Sigma_a)$.

Using the above models, they obtain the benchmarked sub-annual estimates by applying stochastic least squares. That is, they minimize $E(\hat{\theta}_t - \theta_t)^2$, the mean squared error. This technique is also referred to in time series terminology as signal extraction, and the derivation of the solution can be found in the paper written by Hillmer and Trabelsi.

As it is stated with the models, this method takes into account the sampling variances and covariances of the sub-annual and annual estimates. Unfortunately, the approach does not accommodate biases in the sub-annual data. Also, since ARIMA modelling is being used in

this method, it would be costly to implement for large scale surveys dealing with hundreds of series. Therefore it would be best to use this type of approach for only a small number of very important economic indicators. There would also be risks of oversmoothing the data if the ARIMA models are not properly specified.

Cholette and Dagum(1989) modified the Hillmer and Trabelsi approach by introducing an “intervention” model instead of an ARIMA model. This allows the modelling of systematic effects in the time series, but according to the authors, this approach still possesses the same weaknesses as the original Hillmer and Trabelsi method.

3.3 Model on Trends

The following method was developed in an attempt to meet the benchmarking requirements of the economic surveys. It is based on the assumption that the sub-annual estimates follow the model:

$$\frac{y_t}{y_{t-1}} = \frac{\theta_t}{\theta_{t-1}} + v_t \quad t = 1, 2, \dots, n \quad (3.5)$$

and the annual estimates follow model (3.4), where:

$\{\theta_t\}$ is a sequence of sub-annual parameters (true values), as in Denton’s method,

$\{v_t\}$ is a sequence of dependent sub-annual sampling errors of the trends with mean vector and covariance matrix $(\mathbf{0}, \Sigma_v)$.

Least squares theory is applied to the above models to produce benchmarked estimates. The description of the Gauss-Newton algorithm necessary to solve this problem and the calculation of the covariance matrix of the benchmarked estimates are given in Laniel and Fyfe (1989) or (1990).

This method can be used when the benchmarks come from either a census or annual overlapping samples and when the sub-annual level estimates are biased, if the relative bias is a constant. The assumption of a constant relative bias will be verified in practice if the rate of the frame maintenance activities is relatively stable, that is, when the proportion of frame coverage deficiencies is fairly constant over time. This assumption also implies that the under-covered businesses behave like the ones covered by the frame.

One technical problem with this method is that the sampling variance-covariance matrix of the sub-annual trends cannot be calculated directly and an approximation has to be used. The first-order Taylor approximation has been tried but in some cases the resulting sampling variances and covariances were zero or negative when they should be positive. For this reason, an alternative model to (3.5) is presented in the next section.

3.4 Model on Levels

The following method is an alternative to the previous one and is suggested so that the sampling variance-covariance matrix of the sub-annual estimates would be easier to obtain. It assumes that the sub-annual estimates follows the model:

$$y_t = \alpha \theta_t + u_t \quad t = 1, 2, \dots, n, \quad (3.6)$$

where α is a fixed parameter taking into account the constant relative bias and u_t is the same as for equation (3.3). The annual estimates follow model (3.4).

Benchmarked estimates are obtained by applying least squares theory to the above models. The algorithm required to solve this problem is the same as for method 3.3.

3.5 Discussion

Among the methods reviewed here, the most appropriate one for benchmarking a single time series in the context of the large scale surveys is the new approach based on the model on levels. It has a statistical basis which allows us to calculate confidence regions and test the goodness of fit of the benchmarked model. To test for lack of fit one has to be careful in choosing a test since the benchmarked estimates, $\hat{\theta}_t$, have quite a small number of degrees of freedom, $m - 1$ (the number of annual observations minus one), in comparison to the number of observations, $n + m$. This small number of degrees of freedom also suggests that with the model on levels, we can expect to get benchmarked estimates with a chronological pattern similar to the one observed in the sub-annual data.

A current practical issue with benchmarking methods which take into account sampling errors such as in 3.4, is the derivation of sampling covariances between two level estimates corresponding to two different time periods. Should they be calculated directly using the sample design for all pairs of time periods or should they be modelled? From a theoretical point of view, it is better to calculate these directly, since the sequence of sampling errors is intrinsically a non-stationary stochastic process due to the population variance-covariance varying with time. However, calculating all sampling covariances can be cumbersome, thus leaving the issue of how to obtain sampling covariances still an open question.

3.6 An Example

As a comparison between Denton's method described in section 3.1 and the model on the levels approach suggested in section 3.4, these two methods were applied to a special and interesting benchmarking case. It is a situation where the annual estimates have sampling variances six times the size of the sampling variances of the corresponding monthly estimates. In such a case, the advantage of using the model on levels approach instead of Denton's method will be easily observed.

The special case, though possible in practice, was made up of simulated data. Firstly, twenty-four monthly estimates were taken from an existing economic survey. A sampling covariance matrix was arbitrarily given to these monthly estimates. The variances and covariances were calculated in by using an equal coefficient of variation through time and the following correlation pattern:

$$\rho_{ij} = 1 - \frac{|j - i|}{24} \quad \text{for } i = 1, 2, \dots, 24 \quad \text{and } j = 1, 2, \dots, 24$$

where i and j are the indexes of a pair of monthly estimates. Then, two corresponding annual estimates were constructed as follows. The first annual figure was 25% larger than the sum of the first monthly figures. Whereas the second annual figure was only 5% larger than the total of the last twelve monthly observations. The two annual estimates were given sampling variances equal to six times the variances of the corresponding monthly totals and their correlation was fixed at 0.5.

The monthly estimates are represented by the full continuous line and the annual estimates by the horizontal lines on figure 3.1. The two horizontal lines are equal to the values of the

annual figures divided by twelve. On the same figure, the line with long dots represents the monthly series benchmarked with the approach based on the model on levels. The line with short dots is the benchmarked monthly series with Denton's method.

From figure 3.1, it can be observed that the series benchmarked with the model on levels approach has the same year-to-year movement as the original monthly series. Whereas the series benchmarked with Denton's method has the same year-to-year movement as the annual estimates. It can also be seen that both benchmarked series are over the original monthly series.

The difference in the year-to-year movement between the two benchmarked series can be explained as follows. The approach based on the model on levels gives the benchmarked series a year-to-year movement essentially obtained by weighting the annual and sub-annual data with the inverse of their sampling variances. Since, in this example, the sub-annual estimates are much more reliable than the annual estimates, the benchmarked series got the year-to-year movement of the monthly figures. Whereas with Denton's method, the year-to-year movement of the benchmarked series is constrained to one of the annual series regardless of its reliability. In this sense the approach based on the model on levels is better than Denton's method.

As a last comment on this example, the fact that both benchmarked series are above the original monthly series simply illustrates that both methods are providing a correction for the bias of the monthly estimates.

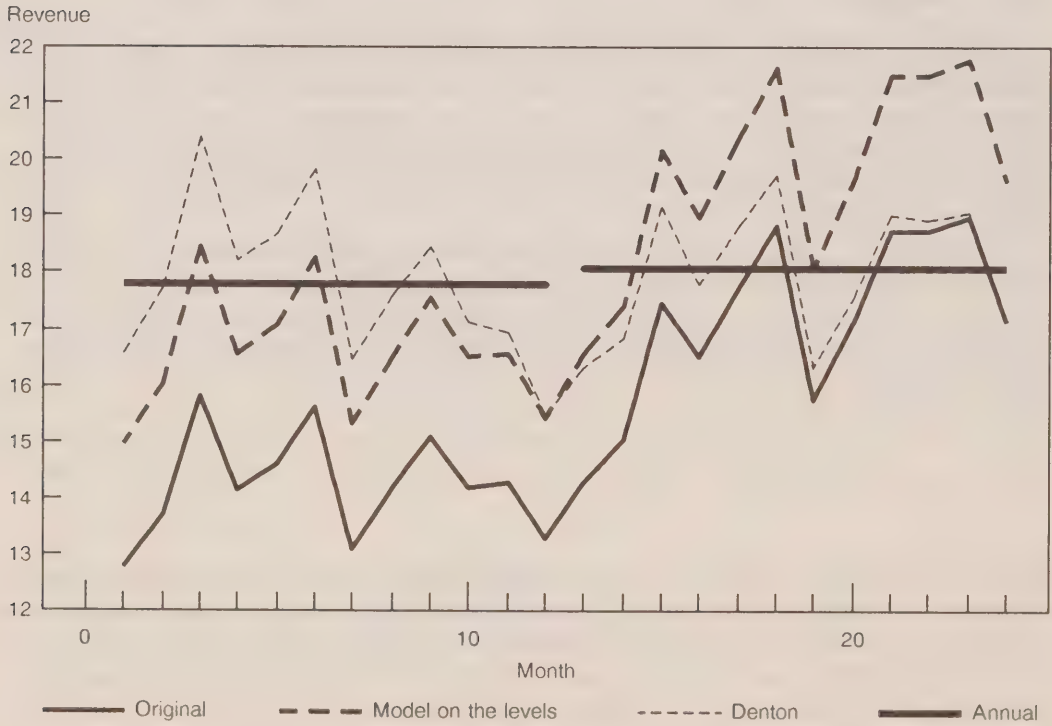


Figure 3.1 Plot of the original and two benchmarked monthly series and of the annual series

4. CONCLUSION

The problem of improving sub-annual survey estimates with the use of annual survey estimates has been examined. A new and simple procedure to benchmark a single time series has been presented. This procedure could be implemented in a computer system to allow its use in an automated mode. The advantage of the procedure over more traditional methods (e.g., Denton's) is that it takes account of sampling errors. Some issues in using the proposed procedure for benchmarking a single time series have been discussed. Two important practical questions have been pointed out: benchmarking a table of series and preliminary benchmarking. Approaches to address these two topics have to be explored.

REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day.
- CHOLETTE, P.A. (1988a). Benchmarking and Interpolation of Time Series. Statistics Canada, Working Paper No. TSRA-87-014E.
- CHOLETTE, P.A. (1988b). Benchmarking Systems of Socio-Economic Time Series. Statistics Canada, Working Paper No. TSRA-88-017E.
- CHOLETTE, P.A., and DAGUM, E.B. (1989). Benchmarking Socio-Economic Time Series Data: A Unified Approach. Working Paper No. TSRA-89-006E, Statistics Canada.
- DENTON, F.T. (1971). Adjustment on Monthly or Quarterly Series to Annual Totals: An approach Based on Quadratic Minimization. *Journal of the American Statistical Association*, 66, 99-102.
- HILLMER, S.C., and TRABELSI, A. (1987). Benchmarking of Economic Time Series. *Journal of the American Statistical Association*, 82, 1604-1071.
- LANIEL, N., and FYFE, K. (1989). Benchmarking of Economic Time Series. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.
- LANIEL, N., and FYFE, K. (1990). Benchmarking of Economic Time Series. Business Survey Methods, Statistics Canada, Business Survey Redesign Project Working Paper.
- MONSOUR, N.J., and TRAGER, M.L. (1979). Revision and Benchmarking of Business Time Series. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

Forgot the Sampling Scheme at the Estimation Stage?

SHIBDAS BANDYOPADHYAY¹

ABSTRACT

For a class of linear unbiased estimators in a class of sampling schemes, it is shown that one can forget the weights used for sample selection while estimating a population ratio by a ratio of two unbiased estimators, respectively of the numerator and the denominator defining the population ratio. This class of schemes includes commonly used sampling schemes such as unequal probability sampling with or without replacement, stratified proportional allocation sampling with unequal selection probabilities and without replacement in each stratum, *etc.*

KEY WORDS: Ratio of unweighted totals; Symmetric sampling.

1. INTRODUCTION

Let m be the number of adult literates among t adult members in a sample of n families drawn from a given population. Let the population adult literacy rate R be estimated as $r = m/t$. Similarly, for a two-way table giving percentage distribution of persons by age-group and sex, let a cell entry be estimated by a ratio (multiplied by 100 to make it a percentage) of the number of persons classified into the cell to the total number of persons, in the sample of n families.

Irrespective of the method of selection of the families, this simple ratio of two unweighted totals for estimating a ratio or a percentage distribution is acceptable to many non-statistical users. Indeed, in some survey reports, tables giving percentage distributions or rates are so computed, as if the sampling scheme had been a self-weighting one.

If, however, the sampling scheme for selecting the n families had been a (single stage) PPSWOR, one is expected to go about finding weighted totals for obtaining unbiased estimators of numerators and respective denominators before computing a ratio or a percentage distribution.

This study shows that, for sampling schemes such as a single stage PPSWOR but without any further assumptions,

- (i) a ratio of two unweighted totals estimates the corresponding population ratio, as a ratio of an *unbiased estimator* of the numerator to an *unbiased estimator* of the respective denominator;
- (ii) there is a class of sampling schemes, other than self-weighting designs, for which (i) holds. This class includes one stage unequal probability, with or without replacement, sampling schemes and stratified proportional allocation sampling with unequal probability without replacement selection in each stratum.

2. SYMMETRIC SAMPLING SCHEMES

Consider a finite population consisting of N units U_1, U_2, \dots, U_N . Let Y_i and X_i , denote the values of two study variables, Y and X respectively, associated with the unit U_i , $i = 1, 2, \dots, N$.

¹ Shibdas Bandyopadhyay; Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

The problem is to estimate a rate or a ratio $R = T(Y)/T(X)$ where $T(Y) = Y_1 + Y_2 + \dots + Y_N$, and $T(X)$ is similarly defined with the variable X .

The usual procedure is to estimate $T(Y)$ and $T(X)$ unbiasedly and take their ratio to estimate R . The aim of this paper is to follow the same procedure in such a way that the ratio becomes free of the selection probabilities of the sample units.

Fix a sampling scheme.

Let S denote the set consisting of all possible samples such that $p(s) > 0$, where $p(s)$ denotes the probability of drawing the sample s , and $\sum_{s \in S} p(s) = 1$.

For s in S and $i = 1, 2, \dots, N$,

$n(i, s)$ = the number of times U_i is included in s , and $\alpha_i = \sum_{s \in S} n(i, s)$, the number of times U_i is included in all possible samples.

$S, p(s), \alpha_i$ depend on the sampling scheme.

Definition 2.1. A sampling scheme is said to be symmetric if $\alpha_i = \alpha$, for all $i = 1, 2, \dots, N$.

The following estimator, based on the sample s , in the class of linear unbiased estimators of Godambe (1955) for $T(Y)$, was studied by Bandyopadhyay *et al.* (1977).

$$T(Y, s) = \sum_{i=1}^N Y_i n(i, s) \alpha_i^{-1} p^{-1}(s). \quad (2.1)$$

Clearly, $T(Y, s)$ is unbiased for $T(Y)$. An estimator of the ratio $R = T(Y)/T(X)$, as a ratio of an unbiased estimator of $T(Y)$ to an unbiased estimator of $T(X)$, based on a sample s , is

$$R(s) = T(Y, s)/T(X, s) = \sum_{i=1}^N Y_i n(i, s) \alpha_i^{-1} \Bigg/ \sum_{i=1}^N X_i n(i, s) \alpha_i^{-1}. \quad (2.2)$$

For symmetric sampling schemes, $\alpha_i = \alpha$ for all i and (2.2) becomes

$$R(s) = \sum_{i=1}^N Y_i n(i, s) \Bigg/ \sum_{i=1}^N X_i n(i, s) =$$

$$\frac{\text{unweighted total of } Y \text{ values in the sample}}{\text{unweighted total of } X \text{ values in the sample}} \quad (2.3)$$

and the above observations are summarized in the following theorem.

Main theorem. For a symmetric sampling scheme, a ratio of two unweighted totals estimates the corresponding population ratio as a ratio of an unbiased estimator of the numerator to an unbiased estimator of the respective denominator, but the estimated ratio does not involve the selection probabilities of the population units in the sample.

It may be noted that the inclusion probabilities of the units in the sample need not be equal for symmetric sampling schemes. Thus, symmetric sampling schemes need not be self-weighting. Self-weighting designs require constancy of $\alpha_i p(s)$ for all i and s , and constancy of $\alpha_i p(s)$ for all i and s does not make the sampling scheme symmetric.

For a non-symmetric scheme, (2.2) is easy to compute as α_i 's are easy to compute in most cases and there is no need to compute inclusion probabilities.

For without replacement sampling of n units, there are $\binom{N-1}{n-1}$ (un-ordered) samples containing a given unit U_i , so $\alpha_i = \binom{N-1}{n-1}$ for all i and thus, in particular, PPSWOR is symmetric. It may be noted that not all PPSWOR schemes result in $\binom{N}{n}$ possible samples. As noted in Connor (1966), in some cases systematic PPS samples in a pre-determined order or randomized PPS systematic sampling may result in zero probability for some set of n units. The result applies if the PPSWOR scheme is such that no joint inclusion probability of any set of n units is zero.

For with replacement sampling of n units, there are N^n (ordered) samples and so $\alpha_i = nN^{n-1}$ for all i and thus, in particular, PPSWR is symmetric.

For PPSWOR in each of k strata, the α -value for each unit in the j th stratum is

$$\alpha_j = \frac{n_j}{N_j} \prod_{i=1}^K \binom{N_i}{n_i}$$

which becomes a constant when allocation is proportional and if no joint probability of any set of units in any stratum is zero, where N_j and n_j are respectively the population and sample sizes for the j th stratum, $j = 1, 2, \dots, k$. Similar allocation may be made to make a multistage sampling scheme symmetric.

For PPSWR sampling, it may be noted that the unbiased estimator of $T(Y)$ given by (2.1) is inadmissible. This estimator can be improved upon by putting $n^*(i,s)$ and α_i^* respectively for $n(i,s)$ and α_i , where $n^*(i,s)$ is 1 if $n(i,s)$ is at least 1 and $n^*(i,s)$ is zero if $n(i,s)$ is zero, and α_i^* is α defined with $n^*(i,s)$. Here, $\alpha_i^* = N^n - (N-1)^n$, the number of (ordered) samples containing a given unit U_i . It has not been possible to obtain a mathematical expression for relative efficiency in a closed form for comparison, even with respect to PPSWR schemes.

Among the possibilities for comparison of relative bias and relative efficiency, an empirical study is included for comparison with PPSWOR scheme. Another attractive possibility is to study large sample variance and bias using Taylor series expansions.

It is clear that it is not possible to estimate the variance of $R(s)$ without the weights or further assumptions. However, if s_1 and s_2 are two half-samples drawn by the same symmetric sampling scheme (like two independent PPSWOR samples of equal size), R is estimated as $[R(s_1) + R(s_2)]/2$, and its unbiased variance estimator is $[R(s_1) - R(s_2)]^2/4$.

If $T(X)$ is known, a ratio-type estimator for $T(Y)$ is $T(X)T(Y,s)/T(X,s)$, which may be improved as in Bandyopadhyay (1980) depending on whether or not the sampling fraction is more than half.

When the population units are divided into k non-overlapping clusters and the selection probability of the j th cluster is p_j then the design become symmetric with $\alpha_i = 1$ for all units in all the clusters. It may be noted that the sample size is the size of the selected cluster and so, the symmetric sampling schemes need not be fixed sample size designs.

3. EMPIRICAL STUDY ON BIAS AND MEAN SQUARE ERROR

Yates and Grundy (1953) considered the following three hypothetical populations, each with 4 population units.

	Population A				Population B				Population C			
X	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Y	0.5	1.2	2.1	3.2	0.8	1.4	1.8	2.0	0.2	0.6	0.9	0.8

The sampling scheme is to draw a sample of size $n = 2$ by PPSWOR using X -values as size measure. It is proposed to compare bias and mean square error of $R(s)$ with those of $R_{HT}^{(s)}$ where $R_{HT}^{(s)}$ is the ratio of the Horvitz-Thompson (1952) estimator of $T(Y)$ to that of $T(X)$. The result of the comparison is presented below.

Populations:	A	B	C
Relative bias of $R(s)$	0.02456	-0.02785	-0.00496
Relative bias of $R_{HT}(s)$	-0.00379	0.00552	0.00232
MSE of $R(s)$	0.2946	0.2946	0.0824
MSE of $R_{HT}(s)$	0.3159	0.3642	0.0690
Relative efficiency of $R(s)$ to $R_{HT}(s)$	1.0723	1.2362	0.8374

Though the absolute bias of $R(s)$ relative to R is more than that of $R_{HT}^{(s)}$ for the three populations, differences are small. $R(s)$ is a more efficient estimator in populations A and B and $R_{HT}(s)$ is more efficient in population C .

Since the above three populations are more extreme than the situations usually met with in practice, it is anticipated that $R(s)$ may be useful when the sampling scheme is not available at the estimation stage.

ACKNOWLEDGEMENT

The author sincerely appreciates active and constructive comments from the referees leading to the final form of this paper.

REFERENCES

BANDYOPADHYAY, S., CHATTOPADHYAY, A.K., and KUNDU, S.C. (1977). On estimation of population total. *Sankhyā*, Ser. C, 39, 28-42.

BANDYOPADHYAY, S. (1980). Improved ratio and product estimators. *Sankhyā*, Ser. C, 42, 45-49.

CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-396.

GODAMBE, V.P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, Ser. B, 17, 269-278.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

YATES, F., and GRUNDY, P.M. (1953). Selections without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Ser. B, 15, 253-261.

Estimation of Panel Correlations for the Canadian Labour Force Survey

HYUNSHIK LEE¹

ABSTRACT

The Canadian Labour Force Survey uses the rotation panel design. Every month, one sixth of the sample rotates and five sixths remain. Hence, under this rotation scheme, once a rotation panel enters in the sample, it stays 6 months in the sample before it rotates out. Because of this design feature and the way of selecting the rotate-in panel, the estimates based on the panels in the same or different months are correlated. The correlation between two panel estimates is called the panel correlation. Three kinds of panel correlations are defined in this paper: (1) the correlation (denoted by ρ) between estimates for the same characteristic based on the same panel in different months; (2) the correlation (denoted by γ) between estimates of the same characteristic based on geographically neighboring panels in different months; (3) the correlation (denoted by τ) between estimates of different characteristics based on the same panel in the same or different months. This paper describes a methodology for estimating these panel correlations and presents estimated correlations for selected variables using 1980-81 and 1985-87 data with some discussion.

KEY WORDS: Repeated panel survey; Rotation; Taylor method.

1. INTRODUCTION

The Labour Force Survey (LFS) is a continuing monthly household survey which employs rotating panel design. The sample consists of six equal size rotation panels one of which is replaced by a new panel each month. The rotated-in panel stays in the sample for six months before it rotates out from the sample. (For detailed description of the LFS methodology, readers are referred to Platek and Singh (1976) and Singh *et al.* (1990).) Therefore, the estimates based on the same panel consisting of the same sampling units in different months are highly correlated. Moreover, an outgoing rotation panel is usually replaced by a neighboring panel. Because they are geographically close, estimates based on these neighboring rotation panels are also correlated. These correlations are called panel correlations. In this paper, we will describe and discuss how the panel correlations can be estimated and present their estimates for selected variables. The work was originated for the study of composite estimation technique. However, the results are applicable in any situation where the panel correlation plays a role.

The paper is structured as follows. In Section 2, necessary definitions, notations and assumptions are given. Methodology is described in Section 3 and results and discussion are given in Section 4.

2. DEFINITIONS OF PANEL CORRELATION COEFFICIENTS

To define various panel correlations we need to define common panels and the predecessor panel. A panel is identified by the panel number which indicates the duration of the panel in the sample. Thus, Panel 1 in month m , becomes Panel 2 in month $m + 1$, Panel 3 in month $m + 2$,

¹ H. Lee, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Table 1
Common and Predecessor Panels Pertaining to Months m and $m - j$

m	$m - 1$	$m - 2$	$m - 3$	$m - 4$	$m - 5$	$m - 6$	$m - 7$	$m - 8$	$m - 9$	$m - 10$	$m - 11$
1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))	((2))
2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))
3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))
4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))
5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))
6	5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)

Note: Single and double parentheses indicate *single* and *double* predecessors, respectively.

and so on. Another term *rotation group* is often used to identify a panel regardless of its duration in the sample. For instance, Rotation Group 1 which rotates in in January is identified as Rotation Group 1 throughout its stay in the sample until it rotates out in July. Then, Panel 1 in January indicates Rotation Group 1 and Panel 2 in February indicates the same rotation group which is now two months old and so on.

Two panels in two different months which represent the same rotation group are called *common panels*. When a rotation group rotates out, it is usually replaced by a rotation group consisting of neighboring households and given the same rotation group number. A panel associated with the out-going rotation group is called a *predecessor panel* of a panel associated with the in-coming rotation group. Therefore, in the above example, Panel 6 in June which is associated with Rotation Group 1 is a predecessor panel of Panel 1 in July. Table 1 shows schematically the common and predecessor panels pertaining to given months m and $m - j$.

Since each panel can be identified by two components, month and panel number, let $P(\text{month}, \text{panel number})$ denote a panel. Then $P(m, 4)$ and $P(m - 1, 3)$, for instance, are common panels 1 month apart. Similarly, $P(m, 4)$ and $P(m - 2, 2)$ are common panels 2 month apart. The correlation coefficient of estimates of a characteristic based on common panels that are j months apart is denoted by ρ_j . Obviously, there are no common panels which are more than 5 months apart and thus, the subscript j can be at most 5. We assume that ρ_j is independent of m and panel number. However, it is a function of j and varies between characteristics.

The correlation coefficient of estimates based on a panel and its predecessor that are j months apart is denoted by γ_j . But in this case, j can go up to 11, *i.e.* γ_{11} is the last correlation coefficient in this series and it is the correlation between $P(m, 6)$ and $P(m - 11, 1)$. We assume again that γ 's are independent of m and panel number. They do, however, depend on characteristic as well as j as ρ -correlations do.

The third type of panel correlation is defined as the correlation between estimates for two different characteristics based on common panels and denoted by τ_j for common panels that are j months apart. Now j can take values from 0 to 5. The same assumptions as for the ρ 's and γ 's apply here as well.

The formal definitions of ρ 's, γ 's and τ 's are as follows:

Let $y_{m,l}$ be the LFS estimate of a characteristic of interest obtained from $P(m,l)$. We assume that $V(y_{m,l}) = \sigma_y^2$ regardless of m and l . Then, ρ_j 's are defined by

$$\text{Cov}(y_{m,l}, y_{m-j,l-j}) = \rho_j \sigma_y^2, \quad 1 \leq j \leq 5, \quad j < l \leq 6,$$

and γ_j 's by

$$\text{Cov}(y_{m,l}, y_{m-j,6+l-j}) = \gamma_j \sigma_y^2,$$

where $1 \leq l \leq j$ if $1 \leq j \leq 6$ and $j - 5 \leq l \leq 6$ if $7 \leq j \leq 11$.

It would be natural to conjecture that ρ_j 's and γ_j 's decrease as the subscript j increases and that ρ_j 's are larger than γ_j 's because ρ_j 's are correlations pertaining common households while γ_j 's are those pertaining neighboring households. We can also define the correlation between a panel and the predecessor of the panel's predecessor (denoted by double parentheses and called *double predecessor* in Table 1) in a similar way, say δ , and thus, we have $\delta_7, \delta_8, \dots, \delta_{17}$. They will be smaller than γ_j 's but could be quite close to them for the same subscript because double and single predecessors are close geographically. However, the δ -correlations are not considered here due to time and resource constraints.

We assume that $\text{Cov}(y_{m,l}, y_{m,l'}) = 0$ if $l \neq l'$ and $\text{Cov}(y_{m,l}, y_{m-j,l'}) = 0$ if $P(m-j, l')$ is not a common panel nor a predecessor of $P(m, l)$.

In order to define τ -correlations, let $x_{m,l}$ be the LFS estimate of another characteristic obtained from $P(m, l)$ and let $V(x_{m,l}) = \sigma_x^2$ be independent of m and l . Then τ -correlations are defined by

$$\text{Cov}(y_{m,l}, y_{m-j,l-j}) = \tau_j \sigma_x \sigma_y, \quad 0 \leq j \leq 5, \quad j < l \leq 6.$$

3. ESTIMATION OF THE PANEL CORRELATIONS

Since a variance estimation computer program was available, the method described here was geared to use this program with minimum modification. The methodology used in the program is the generalized Keyfitz method (Choudhry and Lee 1987; Lee 1989a) better known as the Taylor method. The program can compute variance estimates of linear combinations of monthly estimates.

We employ the following basic equality to estimate the desired correlations using the existing variance program:

$$\text{Cov}(A,B) = \frac{V(A) + V(B) - V(A - B)}{2}. \tag{1}$$

From the program, $V(A - B)$, $V(A)$ and $V(B)$ can be obtained and so can $\text{Cov}(A,B)$ using (1). An expression for $V(A - B)$ from which (1) can be obtained is also given in Kish (1965).

3.1 Estimation of ρ -Correlations

Let $A = \sum_{l=2}^6 y_{m,l}$ and $B = \sum_{l=1}^5 y_{m-1,l}$. A and B are obtained by eliminating Panel 1 from month m and Panel 6 from month $m - 1$, respectively. Note that the eliminated panels are uncommon and the remaining ones are all common. Using the variance program we compute estimates of $V(A - B)$, $V(A)$ and $V(B)$ and obtain estimates of $\text{Cov}(A,B)$ by (1). From the assumptions given in Section 2, it is easy to see that

$$\begin{aligned} \text{Cov}(A,B) &= 5\rho_1\sigma_y^2, \\ V(A) &= V(B) = 5\sigma_y^2, \end{aligned}$$

and thus,

$$\rho_1 = \frac{\text{Cov}(A, B)}{\sqrt{V(A)V(B)}}. \quad (2)$$

An estimate of ρ_1 is then obtained by substituting estimates of $\text{Cov}(A, B)$, $V(A)$ and $V(B)$. Estimates of ρ_2 , ρ_3 and ρ_4 can be obtained in the same way by putting $A = \sum_{l=j+1}^6 y_{m,l}$, and $B = \sum_{l=1}^{6-j} y_{m-j,l}$, $j = 2, 3, 4$. But there is some problem in estimating ρ_5 this way. When we drop all uncommon panels from months m and $m - 5$, only one panel is left in each month and this causes problem in variance estimation for Self-Representing Units (SRUs). SRUs are large cities each of which is represented in the survey by independent sampling. There is no such problem for Non-Self-Representing Units (NSRUs) which are the areas outside of the SRUs, containing rural areas and small urban centers. In NSRUs, each Primary Sampling Unit (PSU), which becomes a replicate for variance estimation, has all rotation panels and thus, even after eliminating 5 uncommon panels, there is still one panel remaining in the PSU so that variance can be computed. In SRUs, however, rotation panels form replicates and if there is only one panel left, then there is only one replicate in each stratum and thus, variance can not be computed in the usual way. Therefore, $\hat{\rho}_5$ was obtained by prediction using a nonlinear regression $\rho = a + bt + ce^{-t}$, $t = 1, \dots, 4$. Another way to estimate ρ_5 will be discussed later in Subsection 4.1.

3.2 Estimation of γ -Correlations

It is easy to see that $\text{Cov}(A, B) = (5\rho_1 + \gamma_1)\sigma_y^2$ if $A = \sum_{l=1}^6 y_{m,l}$ and $B = \sum_{l=1}^6 y_{m-1,l}$. In general,

$$\text{Cov}(A, B) = \{(6 - j)\rho_j + j\gamma_j\}\sigma_y^2,$$

where

$$A = \sum_{l=1}^6 y_{m,l},$$

$$B = \sum_{l=1}^6 y_{m-j,l}, \quad j = 1, \dots, 4.$$

Then, an estimate of γ_j can be obtained from the following equation:

$$\gamma_j = \frac{1}{j} \left[6 \frac{\text{Cov}(A, B)}{\sqrt{V(A)V(B)}} - (6 - j)\rho_j \right], \quad (3)$$

by substituting estimated values on the right. There is a direct way to estimate these γ -correlations including γ_5 by

$$\gamma_j = \frac{\text{Cov}(A_j, B_j)}{\sqrt{V(A_j)V(B_j)}}, \quad (4)$$

where $A_j = \sum_{l=1}^j y_{m,l}$ and $B_j = \sum_{l=7-j}^6 y_{m-j,l}$, $j = 2, \dots, 5$. In Section 4, the two methods were compared by using empirical data.

Other γ -correlations ($\gamma_j, j = 6, \dots, 10$) are obtained by (4) with

$$A_j = \sum_{l=j-5}^6 y_{m,l},$$

$$B_j = \sum_{l=1}^{12-j} y_{m-j,l}.$$

There is no simple way of estimating γ_{11} directly or indirectly. Both $\hat{\gamma}_5$ and $\hat{\gamma}_{11}$ were predicted by a log-linear model $\gamma = \exp(a + bt)$, $t = 1, \dots, 4, 6, \dots, 10$.

3.3 Estimation of τ -Correlations

These correlations can be estimated by the same way as the ρ -correlations just by replacing $y_{m,l}$ by $x_{m,l}$. Let $A = \sum_{l=j+1}^6 x_{m,l}$ and $B = \sum_{l=1}^{6-j} y_{m-j,l}$, $j = 0, 1, \dots, 4$. Then we have

$$\text{Cov}(A,B) = (6 - j) \tau_j \sigma_x \sigma_y,$$

$$V(A) = (6 - j) \sigma_x^2,$$

$$V(B) = (6 - j) \sigma_y^2,$$

from which we get

$$\tau_j = \frac{\text{Cov}(A,B)}{\sqrt{V(A)V(B)}}, \quad j = 0, 1, \dots, 4. \quad (5)$$

All τ 's can be estimated using (5) except τ_5 which is predicted by a log-linear model, $\tau = \exp(a + bt)$, $t = 1, \dots, 4$.

4. RESULTS AND DISCUSSION

By using the methods discussed in the previous section, estimates of ρ - and γ -correlations were computed from the 1980-81 and 1985-87 LFS data for 5 characteristics: In Labour Force (IN LF), Employed (EMP), Employed Agriculture (EMP AG), Employed Non-Agriculture (EMP NON-AG), Unemployed (UNEMP). The panel correlations were estimated for only 3 provinces, Nova Scotia (NS), Ontario (ONT), and British Columbia (BC) from the 1980-81 data. However, the estimation was extended to all provinces when more recent data (March 1985 - February 1987) were used. Moreover, 4 more characteristics, the employed and the unemployed of two age groups, 15-24 and 25+ (EMP 15-24, EMP 25+, UNEMP 15-24, UNEMP 25+), were added. The estimation of τ -correlations was done only for those additional characteristics for NS, ONT and Alberta (ALT) from the 1985-87 data.

In the following, only part of these results will be presented and discussed. All the results are available in Lee (1989b).

4.1 Estimates of ρ -Correlations

The results of estimated ρ -correlations are given in Table 2. Even though estimates for the 5 characteristics (IN LF, EMP, EMP AG, EMP NON-AG, UNEMP) from the 1985-87 data are available for all provinces, the results for only 3 provinces, NS, ONT and BC, are presented for a historical comparison. Table 2 also shows the results for the other 4 characteristics (EMP 15-24, EMP 25 + , UNEMP 15-24, UNEMP 25 +) from the provinces of NS and ONT.

The ρ -correlations are generally high as expected because they are correlations for the common panels. The correlations for EMP AG are the highest and those for UNEMP are the lowest. It seems that the size of the ρ -correlation indicates the degree of mobility of the labour force with a particular characteristic. For instance, the high ρ -correlation for EMP AG shows a low mobility of the labour force in agriculture while a high mobility of unemployed labour force is demonstrated in its low ρ -correlation. The different levels of mobility of labour force in two age groups are also evident. The younger group (15-24) is more mobile than the older one (25 +).

The decreasing trend of the ρ -correlations over time is clearly demonstrated in the results. The trend was extremely well fitted by a nonlinear regression model $\rho_t = a + bt + ce^{-t}$. The R-squares (multiple correlations) are close to 1 (> 0.98). Therefore, the predicted values for ρ_5 seem to be very good. In Lee (1989a and 1989b), $\hat{\rho}_5$ was obtained by extrapolating $\hat{\rho}_3$ and $\hat{\rho}_4$ instead. The differences between the predicted and extrapolated values for $\hat{\rho}_5$, however, are very small. They are less than 0.01 for all characteristics except for UNEMP, UNEMP 15-24 and UNEMP 25 + where the largest difference is 0.03.

Table 2
Estimates of ρ -Correlations (1980-81 and 1985-87 Data)

Prov	Characteristic	80-81 Data					85-87 Data				
		$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
NS	IN LF	0.862	0.797	0.744	0.679	0.622	0.845	0.769	0.730	0.696	0.670
	EMP	0.866	0.783	0.714	0.651	0.590	0.863	0.768	0.713	0.686	0.660
	EMP AG	0.913	0.837	0.756	0.678	0.598	0.912	0.867	0.825	0.802	0.773
	EMP NON-AG	0.865	0.774	0.710	0.649	0.594	0.873	0.779	0.724	0.697	0.670
	UNEMP	0.590	0.455	0.333	0.243	0.145	0.703	0.546	0.426	0.415	0.375
	EMP 15-24						0.773	0.632	0.556	0.495	0.446
	EMP 25 +						0.878	0.800	0.754	0.729	0.705
	UNEMP 15-24						0.618	0.454	0.364	0.300	0.246
	UNEMP 25 +						0.695	0.554	0.443	0.440	0.406
ONT	IN LF	0.843	0.782	0.717	0.674	0.622	0.846	0.781	0.732	0.681	0.635
	EMP	0.852	0.779	0.709	0.664	0.611	0.853	0.771	0.706	0.648	0.592
	EMP AG	0.955	0.926	0.901	0.861	0.827	0.962	0.948	0.944	0.937	0.934
	EMP NON-AG	0.861	0.791	0.724	0.678	0.625	0.866	0.795	0.746	0.701	0.660
	UNEMP	0.580	0.445	0.334	0.286	0.222	0.579	0.436	0.328	0.291	0.238
	EMP 15-24						0.747	0.605	0.500	0.429	0.356
	EMP 25 +						0.888	0.824	0.777	0.732	0.691
	UNEMP 15-24						0.468	0.339	0.257	0.219	0.178
	UNEMP 25 +						0.622	0.468	0.365	0.313	0.256
BC	IN LF	0.849	0.767	0.705	0.665	0.622	0.817	0.753	0.701	0.647	0.597
	EMP	0.835	0.755	0.695	0.651	0.607	0.851	0.770	0.711	0.651	0.597
	EMP AG	0.896	0.809	0.733	0.656	0.582	0.938	0.886	0.847	0.828	0.805
	EMP NON-AG	0.855	0.769	0.715	0.661	0.616	0.857	0.784	0.730	0.679	0.632
	UNEMP	0.516	0.407	0.334	0.320	0.294	0.634	0.524	0.459	0.363	0.290

4.2 Estimates of γ -Correlations

As mentioned in Subsection 3.2, there are two ways of estimating γ_2 , γ_3 and γ_4 , that is, by formulae (3) and (4). We will call the method by (3) as Method 1 and that by (4) as Method 2. Only Method 1 can be used to estimate γ_1 while direct estimation of γ_5 is feasible only by Method 2. The two methods are compared in Table 3 using empirical data. In the table, $\hat{\gamma}_5$'s for Method 1 are predicted values by a log-linear model. The table shows that the two methods produced somewhat different results. The correlations produced by Method 2 clearly show an increasing trend contrary to our intuition while Method 1 gave more acceptable results. Moreover, if we compare these correlations with $\hat{\gamma}_1$ in Table 4A (which had to be estimated by Method 1), Method 1 seems to produce more reasonable results than Method 2. Therefore, we adopted Method 1. However, if everything is correct, the two methods should be equivalent and produce similar results. It seems that the real data do not conform to some extent with the assumptions we made to derive the formulae.

Estimates of the γ -correlations are presented in Tables 4A and 4B. The size of γ -correlations is much smaller than that of ρ -correlations as we expected. But it also reflects differences in mobility of the labour force with different characteristics as seen from the results of ρ -correlations.

The overall trend of $\hat{\gamma}$'s is somewhat fuzzy, especially for the results from the 1985-87 data. There are about 25% of cases – a case is a row entry in the tables – in Table 4B which show an increasing trend. In those cases, the log-linear regression lines have a positive slope even though it is fairly small in magnitude. Moreover, in most of those cases, R-squares are small, which indicates that fittings by the log-linear model are not good. This does not mean, however, that there are other models which can fit the data better. Rather it means that no clear trend is exhibited. Among the cases that show a decreasing trend, about half of the cases have an R-square greater than 0.5.

The results from the 1980-81 data show a quite different picture. There is only one case that shows an increasing trend and most of the cases have R-squares > 0.5 . In fact, the results for NS and BC look more reasonable than those for ONT as far as the trend is concerned.

Table 3
Comparison of Estimates of γ_2 , γ_3 , γ_4 and γ_5 Obtained by Different Methods
(Ontario, 1980-81)

Characteristic	Method	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$
IN LF	1	0.141	0.128	0.133	0.135
	2	0.107	0.105	0.116	0.120
EMP	1	0.136	0.142	0.142	0.147
	2	0.100	0.115	0.126	0.133
EMP AG	1	0.483	0.474	0.486	0.451
	2	0.321	0.370	0.407	0.448
EMP NON-AG	1	0.150	0.147	0.157	0.163
	2	0.117	0.134	0.145	0.149
UNEMP	1	0.074	0.076	0.063	0.080
	2	0.043	0.056	0.046	0.043

Note: Methods 1 and 2 are defined by the formulae (3) and (4) in Section 3, respectively.

Table 4A
Estimates of γ -Correlations
(1980-81 Data)

Prov	Characteristic	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
NS	IN LF	0.288	0.263	0.265	0.250	0.236	0.233	0.211	0.199	0.193	0.167	0.164
	EMP	0.262	0.219	0.228	0.226	0.219	0.239	0.210	0.200	0.188	0.161	0.172
	EMP AG	0.351	0.308	0.283	0.237	0.205	0.190	0.141	0.113	0.063	0.021	0.007
	EMP NON-AG	0.238	0.187	0.189	0.180	0.164	0.151	0.123	0.121	0.136	0.091	0.086
	UNEMP	0.106	0.176	0.091	0.097	0.091	0.076	0.066	0.063	0.066	0.032	0.031
ONT	IN LF	0.161	0.141	0.128	0.133	0.135	0.136	0.125	0.127	0.124	0.122	0.117
	EMP	0.164	0.136	0.142	0.142	0.147	0.149	0.148	0.150	0.153	0.141	0.146
	EMP AG	0.477	0.483	0.474	0.486	0.451	0.474	0.459	0.429	0.394	0.323	0.368
	EMP NON-AG	0.184	0.150	0.147	0.157	0.163	0.167	0.166	0.169	0.174	0.156	0.165
	UNEMP	0.141	0.074	0.076	0.063	0.080	0.051	0.045	0.060	0.077	0.136	0.074
BC	IN LF	0.177	0.137	0.117	0.119	0.119	0.112	0.101	0.112	0.094	0.066	0.070
	EMP	0.211	0.146	0.133	0.107	0.101	0.083	0.050	0.068	0.058	-0.033	-0.015
	EMP AG	0.380	0.311	0.301	0.272	0.241	0.216	0.198	0.170	0.122	0.078	0.071
	EMP NON-AG	0.207	0.166	0.161	0.129	0.108	0.093	0.069	0.038	0.023	-0.004	-0.020
	UNEMP	0.126	0.125	0.114	0.103	0.091	0.076	0.062	0.092	0.032	0.040	0.031

Table 4B
Estimates of γ -Correlations
(1985-87 Data)

Prov	Characteristic	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
NS	IN LF	0.250	0.238	0.247	0.230	0.216	0.204	0.181	0.196	0.189	0.162	0.160
	EMP	0.170	0.183	0.205	0.196	0.185	0.157	0.158	0.194	0.198	0.219	0.198
	EMP AG	0.326	0.296	0.246	0.245	0.265	0.267	0.234	0.217	0.259	0.269	0.231
	EMP NON-AG	0.146	0.168	0.199	0.201	0.178	0.153	0.152	0.189	0.199	0.216	0.201
	UNEMP	0.233	0.267	0.241	0.211	0.206	0.168	0.171	0.176	0.157	0.187	0.147
	EMP 15-24	0.107	0.127	0.140	0.133	0.112	0.105	0.099	0.107	0.090	0.074	0.082
	EMP 25+	0.088	0.075	0.117	0.108	0.100	0.099	0.090	0.103	0.099	0.137	0.118
	UNEMP 15-24	0.051	0.080	0.042	0.024	0.054	0.061	0.079	0.081	0.058	0.011	0.049
	UNEMP 25+	0.155	0.129	0.177	0.171	0.148	0.159	0.158	0.127	0.102	0.134	0.124
ONT	IN LF	0.162	0.138	0.141	0.134	0.132	0.135	0.127	0.116	0.111	0.103	0.101
	EMP	0.114	0.122	0.121	0.122	0.117	0.124	0.119	0.108	0.110	0.112	0.111
	EMP AG	0.508	0.518	0.553	0.561	0.571	0.569	0.582	0.617	0.668	0.650	0.672
	EMP NON-AG	0.133	0.140	0.132	0.140	0.157	0.156	0.168	0.182	0.204	0.205	0.210
	UNEMP	0.030	0.047	0.055	0.047	0.043	0.048	0.039	0.030	0.039	0.048	0.041
	EMP 15-24	0.012	-0.006	0.018	0.031	0.017	0.023	0.011	0.011	0.016	0.044	0.029
	EMP 25+	0.354	0.358	0.349	0.343	0.319	0.312	0.298	0.285	0.276	0.240	0.246
	UNEMP 15-24	0.068	0.039	0.038	0.058	0.033	0.026	0.008	0.018	0.011	-0.002	-0.006
	UNEMP 25+	0.052	0.054	0.033	0.017	0.034	0.033	0.026	0.018	0.021	0.044	0.022
BC	IN LF	0.103	0.095	0.113	0.103	0.090	0.090	0.091	0.083	0.078	0.030	0.055
	EMP	0.125	0.100	0.112	0.111	0.116	0.135	0.123	0.121	0.118	0.095	0.114
	EMP AG	0.394	0.443	0.426	0.401	0.396	0.400	0.401	0.381	0.347	0.334	0.345
	EMP NON-AG	0.080	0.067	0.076	0.072	0.091	0.109	0.111	0.118	0.112	0.106	0.124
	UNEMP	0.096	0.086	0.084	0.080	0.083	0.097	0.068	0.074	0.068	0.083	0.071

Table 5
 Estimates of τ -Correlations
 x_1 : EMP 15-24, x_2 : EMP 25+, x_3 : UNEMP 15-24, x_4 : UNEMP 25+,
 (1985-87 Data)

Province	Characteristic	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
NS	(x_1, x_2)	0.150	0.140	0.148	0.181	0.187	0.196
	(x_1, x_3)	-0.440	-0.275	-0.187	-0.135	-0.039	0.126
	(x_1, x_4)	-0.036	-0.040	-0.043	-0.015	0.024	0.022
	(x_2, x_3)	-0.029	-0.037	-0.078	-0.049	-0.016	-0.038
	(x_2, x_4)	-0.437	-0.374	-0.276	-0.182	-0.231	-0.094
	(x_3, x_4)	0.136	0.127	0.094	0.055	0.049	0.020
ONT	(x_1, x_2)	0.092	0.070	0.055	0.040	0.028	0.010
	(x_1, x_3)	-0.420	-0.267	-0.205	-0.161	-0.145	-0.010
	(x_1, x_4)	-0.065	-0.056	-0.053	-0.036	-0.028	-0.019
	(x_2, x_3)	-0.061	-0.054	-0.054	-0.042	-0.089	-0.074
	(x_2, x_4)	-0.392	-0.303	-0.230	-0.187	-0.181	-0.077
	(x_3, x_4)	0.058	0.043	0.022	0.013	0.022	0.001

4.3 Estimates of τ -Correlations

Table 5 contains estimates of τ -correlations obtained from the 1985-87 data for all possible combinations of EMP 15-24 (denoted by x_1), EMP 25+ (x_2), UNEMP 15-24 (x_3) and UNEMP 25+ (x_4). The correlations between x_1 and x_2 are positive as well as those between x_3 and x_4 . Other correlations are mostly negative. In terms of magnitude, only the correlations pertaining to (x_1, x_3) and (x_2, x_4) are quite different from zero. Others are close to zero. These observations seem to agree with what we understand about the movement of labour force between the employed and the unemployed in the same age group. When the employment increases, the unemployment decreases and vice versa. The trend is obviously upward in these cases.

The data were fit by a log-linear model and τ_5 's were predicted. The model fitting seems reasonable except for the correlations between (x_2, x_3) whose R-squares are very small in both provinces NS and ONT.

4.4 Conclusions

The estimation of correlations from complex survey data is a difficult problem. It is so not because the derivation of formulae is difficult – in fact, the formulae given here are elementary – but because there are many practical constraints in applying the formulae. If we had not made the assumptions in Section 3, the estimation of the panel correlations by using the existing computer program would have been impossible. On the other hand, these assumptions should be conformable to the real data to which the formulae are applied. In our case, there seem to be some unconformable elements in the assumptions we made to the real data, which was indicated by the discrepancy in the results obtained by formulae (3) and (4) (see Table 3). Nevertheless, the estimates are not thought to be unreasonable.

In a study of the composite estimator for the LFS, the results given in this paper were successfully used to compare various composite estimators (Kumar and Lee 1983). Recently

Binder and Dick (1990) proposed a method for analyzing Seasonal ARIMA models by taking the survey errors into account. They applied their technique to the LFS data using the estimated panel correlations. However, in cases when the results to be obtained by the use of the estimated panel correlations are sensitive to the accuracy of these estimates, the results should be interpreted carefully.

ACKNOWLEDGEMENT

The author would like to thank the anonymous referees, Editor M.P. Singh, and Assistant Editor L. Mach for their helpful comments. The earlier version of this paper was benefited by comments of S. Kumar and Y. Bélanger of Statistics Canada.

REFERENCES

- CHOUDHRY, G.H., and LEE, H. (1987). Variance estimation for the Canadian Labour Force Survey. *Survey Methodology*, 13, 147-161.
- BINDER, D.A., and DICK, J.P. (1990). Analysis of seasonal ARIMA models from survey data. *Survey Methodology*, this issue.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.
- LEE, H. (1989a). Variance estimation methodology and general purpose variance estimation system for the Labour Force Survey. Methodology Working Paper Series, SSMD-89-022 E, Statistics Canada.
- LEE, H. (1989b). Estimation of panel correlations for the Canadian Labour Force Survey. Methodology Working Paper Series, SSMD-89-023 E, Statistics Canada.
- PLATEK, R., and SINGH, M.P. (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.
- SINGH, M.P., DREW, J.D., GAMBINO, J., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

First Wave Effects in the U.S. Consumer Expenditure Interview Survey

ADRIANA R. SILBERSTEIN¹

ABSTRACT

Panel responses to the U.S. Consumer Expenditure Interview Survey are compared, to assess the magnitude of telescoping in the unbounded first wave. Analysis of selected expense categories confirms other studies' findings that telescoping can be considerable in unbounded interviews and tends to vary by type of expense. In addition, estimates from the first wave are found to be greater than estimates derived from subsequent waves, even after telescoping effects are deducted, and much of these effects can be attributed to the shorter recall period in the first wave of this survey.

KEY WORDS: Bounding; Telescoping; Recall Bias; Conditioning.

1. INTRODUCTION

Respondents to retrospective surveys are asked to recall details of events within a specific time interval, or reference period, and this task of identifying the correct time in which events occurred may be as difficult as remembering the events. Misdating, or "telescoping", is widely recognized as a source of error in surveys, although it is rarely studied directly (Neter and Waksberg 1965). Respondents tend to include in the report events that occurred outside the reference period (external telescoping), *e.g.*, when events are recalled as more recent than they actually are (forward telescoping). Data that can be validated with independent records show that both forward and backward misdating errors are made by respondents (Mathiowetz 1985). This could be "due to the respondent's wish to perform the task required.... When in doubt, the respondent prefers to give too much information rather than too little" (Sudman and Bradburn 1974, p. 69). The net effect of telescoping is generally forward. Bounding methods are designed to create boundaries around the reference period of the survey report, and, in so doing, avoid misdating errors by respondents. A method for bounding the starting point of the reference period, best applied during the interview, involves comparing events reported in a prior interview and deleting duplicate reports. Extending the reference period up to the interview day is a method commonly used to bound the end of the reference period. "Unbounded" reports result by necessity from one-time surveys, and for questions asked only once or for the first time in panel surveys, since no prior data exist to check for erroneous inclusions. These effects can be reduced by including "anchoring" techniques during the interview, *e.g.* constructing a time line (Mingay 1987, p. 132).

This paper is concerned with reporting levels experienced by first time respondents of panel surveys, and provides a comparative analysis of first and subsequent interview waves. The study investigates potential telescoping, conditioning, and recall length effects in estimates of household expenditures, based on data reported in the U.S. Consumer Expenditure (CE) Interview Survey for the year 1984. This survey is one of two independent components designed to collect national data on household expenditures, the other component being the Diary Survey.

¹ Adriana R. Silberstein is a Mathematical Statistician, Office of Prices and Living Conditions, Statistical Methods Division, U.S. Bureau of Labor Statistics, Washington, D.C. 20212, USA.

The survey is conducted by the Census Bureau under contract to the Bureau of Labor Statistics. The first wave of the CE Interview Survey is used to establish cooperation, collect initial inventory data on household possessions, and bound the second wave. There are four subsequent waves of interviews three months apart, collecting data for the previous three calendar months up to the interview day. The bounding method is as follows. Expenses reported for the portion of the calendar month in which the interview takes place (or "current month") are later transcribed onto the next wave questionnaire; this information is available to the interviewer to check for duplicate reports, but is not read to respondents. Data collected during the first wave pertain to expenditures for the current month and for one previous calendar month; these latter expenditures are excluded entirely from the estimates, while current month expenditures become part of the second wave. More details on collection and estimation methods can be found in the 1984 Bulletin (U.S. Bureau of Labor Statistics 1986), and are discussed by Silberstein and Jacobs (1989).

The findings underscore the need for bounding methods in retrospective data collection, since sizable telescoping effects may be present in unbounded recall. In addition, the analysis points out that first time responses may yield higher estimates even after telescoping effects are deducted. These first wave effects may be a direct result of the shorter recall in this wave of the CE Interview Survey, although other factors are not excluded. A discussion of the analysis used to identify telescoping effects is included in section 2, and estimates of telescoping and first wave effects are included in section 3. Conclusions can be found in section 4.

2. IDENTIFYING TELESOPING EFFECTS

2.1 Method of Analysis

One approach for identifying telescoping errors, discussed by Kalton *et al.* (1989, p. 257), is to examine whether there are duplicates in individual responses to consecutive waves. This micro-level approach is not necessarily accurate, as the respondent for a given household may change from one wave to the next. The method is also impractical, since independent records, needed to reconcile discrepancies on dates, may not be readily available. Duplicate responses may not be recorded as such in an ongoing survey, even when they are identified during the interview, as in the CE Interview Survey. More commonly, telescoping effects are evaluated at the aggregate level, by comparing estimates of unbounded and bounded responses, with certain precautions. Tracking the experience of several panels is advisable in order to overcome seasonal incomparabilities, since bounded responses are reported subsequently to unbounded responses and, therefore, do not refer to the same time interval. Another factor to account for in the comparisons is panel conditioning, a phenomenon that refers to changes in respondent behavior as a result of being part of a panel, or to changes in the quality of reports. The assumptions made and the method of estimation used in this study are discussed in section 3, whereas the preliminary testing procedure is described here.

The first step in the analysis is to ascertain whether symptoms of external telescoping can be detected from the survey data. A level of reporting in the first wave that is higher than expected is an indication of telescoping. Unbounded interviews are known to yield higher estimates than bounded interviews, as documented in several studies that compared unbounded and bounded responses (Neter and Waksberg 1964 and 1965; Murphy and Cowan 1976; Cantor 1985). Another indication is the presence of differential effects across separate types of the collected data. Major sources of differences in the way events are retrieved and stated by respondents are recall bias and telescoping. The relationship of these factors suggests that

smaller expenses are forgotten as time increases, but larger more salient expenses, that tend to be remembered better, are more often telescoped.

Telescoping errors can also occur in bounded responses, causing the forward shifting of data within the reference period (internal telescoping). While overall estimates do not change as a result of these effects, the distribution for the three recall months is affected. Reports of apparel and home furnishing and equipment expenses were selected for the study, because characteristics of these expenses were helpful in the analysis. These commodities include expenditures of various degree of salience, and were grouped accordingly. They also tend to differ by degree of underreporting. Many apparel estimates are 40% below the estimates from the National Accounts (NA), and several estimates for home furnishings and equipment are also lower than NA estimates. Estimates for furniture and selected equipment categories, on the other hand, are only 7% below the independent estimates (Gieseman 1987, p. 11), and higher reports in the first wave can be interpreted as the result of external telescoping.

The hypothesis evaluated is whether the first recall month of bounded waves, *i.e.*, the month prior to the interview, is reported similarly to the past month in the first wave. The Hotelling T^2 was used to test differences in eight expenditure groups within each of the two commodities. Given two vectors of means in a repeated-measures design, a two-tailed .05-level test of $H_0: C\mu = 0$ (equality of means) versus $H_1: C\mu \neq 0$ was applied. H_0 was rejected if:

$$[(C\bar{x})'(CS C')^{-1}C\bar{x}]/[np/(n - (p - 1))] > F_{p, n-p+1}(.05), \quad (1)$$

where \bar{x} is a vector of sample means within each commodity (ordered as shown in the tables), S is the covariance matrix computed with the method of balanced repeated replication ($n = 20$ replicates), C is the contrast matrix shown below, and p is the number of contrasts in C .

$$C_{(px2p)} = \left[\begin{array}{cccc|cccc} 1 & 0 & .. & 0 & -1 & 0 & .. & 0 \\ 0 & 1 & .. & 0 & 0 & -1 & .. & 0 \\ . & . & .. & . & . & . & .. & . \\ . & . & .. & . & . & . & .. & . \\ 0 & 0 & .. & 1 & 0 & 0 & .. & -1 \end{array} \right].$$

Simultaneous confidence intervals for individual comparisons by group were derived using the Bonferroni method (Johnson and Wichern 1988), with percentile $t_n(.05/2p)$. Expenditure means were computed using a log transformation of individual expenses reported in the first recall month. Sample weights included adjustments for nonresponse and subsampling, but excluded final weight factors for population controls, which were not available for the first wave. Note that weight adjustments for the first wave were computed only as part of this research, since they are not needed in the ongoing estimation process.

Data from waves 2 to 5 were combined, since differences between these waves were very small. Responses by participants in all five waves (3200 respondents) were selected to assure comparability between the waves and bounding of waves 2 to 5. Unbounded interviews are experienced by new panel respondents, *e.g.* new occupants at a sample address, and by respondents who do not participate in one or more wave during the panel. In 1984, 89% of the interviews in waves 2 to 5 were bounded, 8% were unbounded because respondents were new to the panel and 3% were unbounded resulting from a previous refusal or other non-cooperation (Silberstein 1988). Estimates are affected by unbounded responses, as pointed out by Biderman and Cantor (1984), but this aspect is not treated directly in this study.

2.2 Test Results

Comparisons between means are shown in Table 1 in the original scale, *i.e.*, without the log transformation used in the statistical tests. The first wave displays higher means in nearly all expense groups, and the overall test is significant. The tests for the individual groups reveal that significant differences are found only for large expenditures, such as coats and jackets in apparel and appliances and furniture in home furnishings and equipment. The groups with significant differences are more represented in wave 1 than in other waves, not surprisingly: they account for 19% of total apparel and 72% of total home furnishings in the first wave, compared to 16% and 67%, respectively, in the first recall month of other waves, as shown in Table 2 (columns 1 and 2). A greater number of expenses are also reported in wave 1 for these groups of expenses (Table 2, columns 3 and 4). In addition, the average dollar value of reported expenses in wave 1 tends to be different from the other waves for big-ticket items (*e.g.*, major appliances), but very similar for smaller items (Table 2, columns 5 and 6).

Table 1
Percent Difference in Expenditure Means

	Wave 1 Versus First Recall Month of Waves 2 to 5	
	% Difference (a)	s
APPAREL: (b)	14.5*	4.9
Coats, jackets, furs, suits	39.6*	12.9
Trousers, slacks, jeans	13.6	9.5
Shirts, blouses, tops	9.7	5.6
Sweaters, dresses, skirts	16.4	4.7
Undergarments, hosiery	6.9	5.4
Miscellaneous and combined clothing	-2.5	7.3
Footwear	2.1	6.1
Other apparel items and services	27.4	25.4
Overall test value:	4.16*	
HOME FURNISHINGS AND EQUIPMENT: (b)	48.6*	8.4
Major appliances	76.1*	27.5
Other appliances	56.3*	17.0
Furniture	111.0*	24.8
Large household and entertainment equipment	34.2*	16.0
Other household and entertainment equipment	19.1*	7.1
Home furnishing repair and services	7.0	14.6
Dishes, decorative items, linens	14.0	16.0
Floor and window coverings	52.5	24.3
Overall test value:	13.86*	

(a) Positive values indicate first wave mean is greater. Base of percentages is mean of first recall month in waves 2 to 5.
(b) Commodity totals not included in overall test.
s Standard error of percent difference.
* Significant ($\alpha = .05$).

Table 2
Comparisons of First Wave and First Recall Month of Subsequent Waves

	Percent of Total Expenses		Percent of Total Number of Expenses		Average Dollar Value of Expenses	
	Wave 1	Waves 2 to 5	Wave 1	Waves 2 to 5	Wave 1	Waves 2 to 5
	(1)	(2)	(3)	(4)	(5)	(6)
APPAREL:	100.0	100.0	100.0	100.0	\$ 35	\$ 33
Coats, jackets, furs, suits	19.2	15.7	9.3	8.6	71	59
Trousers, slacks, jeans	10.7	10.8	10.6	9.8	36	35
Shirts, blouses, tops	10.0	10.4	12.0	12.2	31	29
Sweaters, dresses, skirts	14.3	14.0	13.0	12.4	38	37
Undergarments, hosiery	5.2	5.6	16.8	16.7	11	11
Miscellaneous and combined clothing	15.5	18.2	15.4	16.4	36	38
Footwear	11.7	13.1	12.8	13.6	33	31
Other items and services	13.5	12.2	10.1	10.4	45	40
HOME FURNISHINGS AND EQUIPMENT:	100.0	100.0	100.0	100.0	\$123	\$ 92
Major appliances	11.4	9.6	4.2	3.4	370	277
Other appliances	2.3	2.2	9.2	7.1	29	30
Furniture	28.3	19.9	8.9	7.5	385	251
Large household and entertainment equipment	19.7	21.8	8.8	7.6	262	266
Other household and entertainment equipment	10.7	13.4	22.7	22.8	58	56
Home furnishing repair and services	4.7	6.6	8.4	9.5	67	65
Dishes, decorative items, linens	12.9	16.8	33.1	37.5	46	39
Floor and window coverings	10.0	9.8	4.6	4.5	294	172

These differences can be interpreted in several ways, *e.g.*, they may indicate that more expensive purchases are reported in the first wave, or that purchases reported in the first wave are remembered as more expensive. Another interpretation is that a period of time longer than a month may be covered by respondents when the recall is unbounded, especially for large, easily remembered, expenses. In Table 3, comparisons by wave are extended to include the three recall months of subsequent waves. The findings are consistent with the previous tests, but tend to narrow in on the issue of telescoping effects. These comparisons are made on the basis of reporting rates according to the dollar value of the expense. The reporting rate is defined as the percentage of respondents reporting one or more expense of a given type. Note that individual expenses are generally entered on the questionnaire, with the exception of expenses for the same item, month and person in the family, which are usually reported as combined totals and counted as one "expense".

Table 3
Monthly Reporting Rates by Expense Size

	Wave 1	Waves 2 to 5 by Recall Month		
		First	Second	Third
	Percent of respondents			
	(1)	(2)	(3)	(4)
APPAREL:				
No Apparel Expenses (a)	28.8	29.3	38.2	45.5
Less than \$10	38.4	37.7	27.9	25.4
\$ 10 to \$ 40	57.9	55.2	45.3	41.0
\$ 40 to \$100	35.1*	31.0	26.5	21.0
\$100 and over	17.0*	13.7	11.5	8.8
Wave 1 vs 1st recall month of waves 2 to 5				
Overall test value: 29.1*				
HOME FURNISHINGS AND EQUIPMENT:				
No Home Furnishing Expenses (a)	48.1*	51.2	58.5	62.4
Less than \$10	12.3	12.5	7.5	7.5
\$ 10 to \$ 40	30.9	30.0	25.0	22.1
\$ 40 to \$100	21.3*	18.4	14.9	12.8
\$100 to \$400	18.7*	13.8	12.1	10.3
\$400 and over	8.6*	5.6	5.1	4.6
Wave 1 vs 1st recall month of waves 2 to 5				
Overall test value: 17.0*				

(a) Category included in overall test.

* Significant ($\alpha = .05$).

Consistent with the previous comparisons, the overall test is significant and the individual comparisons show that significantly more respondents report expenses of \$100 or more in the first wave; reporting rates for smaller expenses are not significantly different, instead. When the three recall months are examined, the reporting rates for the first recall month appear to be closer to the first wave than to the other two months. The three recall months in waves 2 to 5 show a familiar pattern of decreased reporting, and noteworthy is the increase in the percent of respondents reporting "no expenses". This pattern is evident in each panel wave, as documented by Silberstein and Jacobs (1989) and further studied by Silberstein (1989), and is more likely due to recall effects than telescoping. When reporting rates are recomputed to include only respondents that report the commodity, it is found there are more similarities among the three recall months in subsequent waves than with the first wave. (The rates can be derived from Table 3, by using the percentage of reporters with expenses as the base.) These reporting rates for home furnishing items of \$100 and over are 53% in the first wave and 40%, 41%, and 40%, respectively, in the three recall months of other waves. For apparel items of \$100 and over the rates are 24% in the first wave and 19%, 19%, and 16%, respectively, in the three recall months of other waves. These differences are believed to be symptomatic of external telescoping in the unbounded recall.

3. ESTIMATING TELESCOPING AND FIRST WAVE EFFECTS

3.1 Telescoping Effects

The hypothesis of equality of means implied the response task in the first wave is similar to the one experienced for the first recall month in subsequent waves. The data did not support the hypothesis, since differential effects were found, suggesting external telescoping in the first wave. The results tend to agree with the notion, forwarded by Loftus (1986, p. 196), that internal telescoping may "arise from a different cognitive mechanism" than external telescoping. A general definition of external telescoping (β), on a monthly basis and assuming no panel conditioning, is given by the ratio of unbounded one month recall (with sample mean \bar{x}_U) and bounded one month recall (with sample mean \bar{x}_B):

$$\beta = (E\bar{x}_U/E\bar{x}_B) - 1. \quad (2)$$

This expression may be an overstatement since conditioning effects contribute to lower values for the bounded mean. Panel responses commonly display a downward trend, due to decreased reporting with increasing time-in-sample (TIS) (Bailar 1989). Conditioning effects (α) between two consecutive waves can be defined by the ratio of the two responses (with sample means \bar{x}_i and \bar{x}_{i+1}):

$$\alpha = 1 - (E\bar{x}_{i+1}/E\bar{x}_i). \quad (3)$$

A number of assumptions were made to develop telescoping estimates from the survey data. Expenditure means of bounded one month recall, needed for comparisons with the first wave, cannot be obtained directly from the three month recall. Monthly means computed by dividing the bounded three month recall by a factor of three are not acceptable, considering the recall loss evident in the third recall month of the CE Interview Survey. As an alternative, the first and second recall months were used to estimate bounded monthly means, assuming that recall bias in the second month is moderate and telescoping into the first recall month is mostly from the second recall month. The estimating method is an adaptation of the model developed by Neter and Waksberg in analyzing the 1960 experimental study of expenditures for Residential Alterations and Repairs (Neter and Waksberg 1964 and 1965). The model implies that telescoping and conditioning effects are multiplicative and conditioning compounds with time-in-sample. Since conditioning effects are derived from relationships observed between second and third waves, two terms are necessary when estimating (2) under the assumption of conditioning. An estimate of telescoping is therefore:

$$b_C = (\bar{x}_U/\bar{x}_B)(1 - a)(1 - a/2) - 1. \quad (4)$$

The derivation of (4) is given in the appendix. The conditioning rate (a) was assumed to be constant between waves, considering the special subset of respondents in all five waves. (The Neter/Waksberg model assumed greater effects between the first and second wave.) Time-in-sample effects appear to be small in the CE Interview Survey, judging from a study that compared responses in waves 2 to 5 (Silberstein and Jacobs 1989). An explanation for this may be that declines in reporting are offset by improvements in reporting, as respondents become more knowledgeable about the reporting process. Two conditioning assumptions provided two estimates of telescoping effects, using (4): $a = 0$ (no conditioning), and $a > 0$ conditioning, equal to the rate observed between second and third waves. Four apparel groups and three home furnishing and equipment groups showed some decline from second to third waves, displayed as positive proportions in column 5 of Table 4. These ratios, while not

Table 4
Telescoping Estimates Based on Expenses

	Telescoping effects b_c				TIS effects
	If $a = 0$		If $a > 0$		a
	%	s	%	s	
	(1)	(2)	(3)	(4)	(5)
APPAREL:	28.4	7.0	—	—	-0.02
Coats, jackets, furs, suits	46.2	14.2	—	—	-0.01
Trousers, slacks, jeans	30.3	8.6	12.3	11.8	0.10
Shirts, blouses, tops	27.7	7.8	17.6	16.7	0.05
Sweaters, dresses, skirts	28.3	5.9	8.7	15.0	0.11
Undergarments, hosiery	22.2	6.9	7.2	12.7	0.08
Miscellaneous and combined clothing	5.2	9.5	—	—	-0.18
Footwear	18.1	7.1	—	—	-0.08
Other items and services	54.9	35.8	—	—	-0.15
HOME FURNISHINGS AND EQUIPMENT:	63.1	8.9	—	—	-0.04
Major appliances	95.4	30.7	—	—	-0.03
Other appliances	76.4	16.1	36.0	19.7	0.16
Furniture	113.3	25.2	—	—	-0.05
Large household and entertainment equipment	38.7	13.1	36.5	33.7	0.01
Other household and entertainment equipment	26.2	8.9	—	—	-0.11
Home furnishing repair and services	15.6	14.5	—	—	-0.29
Dishes, decorative items, linens	45.4	14.4	—	—	-0.06
Floor and window coverings	89.4	38.0	66.8	68.7	0.08

a Time-in-sample (TIS), or conditioning, effects when positive.

s Standard error of percent difference.

significant (.05 level), were applied as the conditioning loss between the first and the second wave. Net increases in reports were not considered realistic for the unknown conditioning between these two waves.

The results give indications of the increase that would occur in the estimates in the absence of bounding. Table 4 shows estimates of telescoping effects in percentage form, excluding conditioning effects (column 1), and including them (column 3). Telescoping levels of 40% or higher are estimated for "Coats, *etc.*" and "Other items and services" (a group that includes watches and jewelry), but much lower levels are estimated for other apparel groups. High telescoping levels (63%, on average) are estimated for home furnishing and equipment expenses. Telescoping estimates decrease considerably when some conditioning effects are taken into account, and would be even lower if greater conditioning effects were assumed between wave 1 and wave 2. While these estimates are affected by sampling variability and the assumptions made, the results are consistent with findings reported in other surveys. Neter and Waksberg (1965) reported average telescoping effects of 55% with no conditioning losses and 39% with conditioning losses, for home improvement expenditures; telescoping effects were much lower for small jobs. Telescoping effects derived from the 1974/75 Crime Survey indicated telescoping effects of 44% for personal victimization incidents and 40% for property victimization (Murphy *et al.* 1976).

3.2 First Wave Effects

Differences in responses between first and subsequent waves reflect many cognitive aspects of panel interviews. This section discusses some of the factors involved, and includes a preliminary investigation of net effects. Provided that respondents participate in the whole panel, there is a progressive relationship between respondent and interviewer and more clear expectations on both sides. Quite a few interview conditions change, however. While in some panel surveys subsequent waves may be presented as follow-ups to the first wave, in the CE Interview Survey respondents are asked to report for a period of time three times as long after the first wave and detailed income information is asked in waves 2 and 5. This greater reporting load, and a resulting faster interview pace, has a negative impact on reporting levels, even for the first recall month of these waves. More expense records, *e.g.*, check books and bills, may be used in these waves compared to the first wave, making the bounded reports less likely to be affected by telescoping within the three recall months. The first wave is an easier interview, especially with regard to categories of expenses sensitive to the length of the reference period and the number of persons in the household, *e.g.* apparel expenses. The relative importance of these factors should be researched in field and laboratory studies.

Separate estimates of first wave means, net of telescoping, were developed using the two sets of telescoping effects shown in Table 4. These means (\bar{x}_{B1}) were derived by dividing the unbounded means by the telescoping estimates:

$$\bar{x}_{B1} = \bar{x}_U / (1 + b_C). \quad (5)$$

Results are summarized by commodity in Table 5. Both estimates of net first wave means are higher than means of waves 2 to 5 for all recall months combined, shown in column 2. The total apparel mean is 10% higher in the first wave when conditioning effects are not included, and 16% higher when they are included. The home furnishing and equipment means are also higher, but at a smaller scale: 3% without conditioning and 5% with conditioning. These estimated effects, remaining after telescoping, are interpreted as resulting from the shorter recall period and lesser reporting load in the first wave. The differences between the two commodities and the results for specific groups of expenditures imply that potential gains in reporting tend to increase for smaller expenses, but become quite marginal for big-ticket items.

Table 5
Summary Comparisons of First Wave and Subsequent Waves
Annual Expenditure Means (Standard errors)

	Wave 1	Waves 2 to 5 All Recall Months (a)	Waves 2 to 5 First recall Month	Wave 1 Net of Telescoping	
				Assuming no TIS Effects	Assuming TIS Effects
	(1)	(2)	(3)	(4)	(5)
APPAREL	\$1,663 (59.6)	\$1,182 (61.7)	\$1,452 (71.0)	\$1,295 (66.2)	\$1,370 (n.a.)
HOME FURNISHINGS AND EQUIPMENT	\$1,972 (85.0)	\$1,179 (59.7)	\$1,327 (73.1)	\$1,209 (61.5)	\$1,235 (n.a.)

(a) Means differ from published 1984 estimates, due to special subset of respondents and missing final weight factors.

4. CONCLUSIONS

This paper provides an investigation of potential telescoping effects in unbounded interviews. These effects appear to be considerable, especially for more salient or prominent events. Results from the U.S. Consumer Expenditure Interview Survey indicate that estimates of large infrequent expenses, based on unbounded one month recall, may be between 30% and 50% overstated. Lower overstatement levels are more likely in estimates of small frequent expenses. These findings are in close agreement with other studies on the subject. The study demonstrates that external telescoping effects are much greater than internal telescoping effects within a three month recall period of subsequent waves. In addition, the first wave of the panel survey studied was found to exhibit higher means than the overall means for subsequent waves, even after estimated telescoping effects were deducted. Since the first wave in this survey has one month recall, it is concluded that considerable improvements in reporting levels can be expected from a shorter recall. The potential gains are estimated to be at least 10% for frequent expenditures, but would become marginal as the value of the expenditure increases.

Although the one month recall is viewed as the major reason for the higher estimates, other factors are not excluded. Conditioning effects, assumed constant in this study, may vary between waves. Estimates of one month recall would be even greater, if higher conditioning effects were assumed between the first and second waves. Cognitive aspects of the interview, *e.g.*, respondents cooperation and involvement, and interviewers' approach to collecting data, should be researched in order to understand panel conditioning. The issue of differential effects by type of expenditure should also be addressed within this context. Field and laboratory studies of these data collection aspects would have implications for improving panel survey methodology.

ACKNOWLEDGMENTS

The author thanks the referees, Stuart Scott, and Sylvia Leaver for their comments.

APPENDIX

(1) Explanation of Selected Expenditure Groups

SELECTED APPAREL

Miscellaneous and combined clothing: nightwear, loungewear, accessories, uniforms, and clothing items for infants under 2.

Other apparel items and services: watches, jewelry, sewing materials for making clothes, repair and alteration services, and clothing rental or storage.

SELECTED HOME FURNISHINGS AND EQUIPMENT

Other appliances: small electric kitchen and personal care appliances.

Large household and entertainment equipment: lawn mowers, window air conditioners, televisions, sound equipment, and bicycles.

Other household and entertainment equipment: radios, tape recorders, tools, calculators, camping or sports equipment, and infants equipment.

(2) Estimates of Telescoping Effects

(Adapted from: Neter and Waksberg (1965), 33-37).

For each expenditure group

Let: \bar{x}_U = unbounded one month recall sample mean;

\bar{x}_B = bounded one month recall sample mean, not directly observed in the CE Interview Survey;

\bar{x}_2, \bar{x}_3 = one-month-average sample means from waves 2 and 3, respectively, computed using first and second recall months.

Define: Telescoping effect β , assuming no conditioning

$$\beta = (E\bar{x}_U/E\bar{x}_B) - 1. \quad (1)$$

Conditioning effect, α , between two consecutive waves

$$\alpha = 1 - (E\bar{x}_{i+1}/E\bar{x}_i). \quad (2)$$

Then, assuming telescoping compounds on conditioning,

$$\beta_C = (E\bar{x}_U/E\bar{x}_B) (1 - \alpha) - 1 \quad (3)$$

is the telescoping effect under conditioning.

Using the estimated conditioning effect between 2nd and 3rd waves, $a = 1 - (\bar{x}_3/\bar{x}_2)$, the estimated mean for bounded one month recall is:

$$\begin{aligned} \bar{x}_B &= (\bar{x}_2 + \bar{x}_3)/2 \\ &= (\bar{x}_2 + \bar{x}_2(1 - a))/2 \\ &= \bar{x}_2(1 - a/2). \end{aligned} \quad (4)$$

Assuming a constant rate of conditioning and using (3) and (4), an estimate of the telescoping effect under conditioning, b_C , is:

$$b_C = (\bar{x}_U/\bar{x}_B) (1 - a) (1 - a/2) - 1. \quad (5)$$

REFERENCES

- BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 1-24.
- BIDERMAN, A.D., and CANTOR, D. (1984). A longitudinal analysis of bounding, respondent conditioning, and mobility as sources of panel bias in the national crime survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 708-713.
- CANTOR, D. (1985). Operational and substantive differences in changing the NCS reference period. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.
- GIESEMAN, R. (1987). The consumer expenditure survey: Quality control by comparative analysis. *Monthly Labor Review*, March, 8-14.
- JOHNSON, R.A., and WICHERN, D.W. (1988). *Applied Multivariate Statistical Analysis*. 2nd Edition, Englewood Cliffs, New Jersey: Prentice Hall, 188-190.
- KALTON, G., KASPRZYK, D., and McMILLEN, D. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 249-270.
- LOFTUS, E. (1986). Survey remembering. *Proceedings of the Second Annual Research Conference, Bureau of the Census, Washington, D.C.*, 193-207.
- MATHIOWETZ, N.A. (1985). The problem of omissions and telescoping error: New evidence from a study of unemployment. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 482-487.
- MINGAY, D.J. (1987). Report on the consumer expenditure survey. *Questionnaire Design: Report on the 1987 BLS Advisory Conference*, (Eds. J. Bienias, C. Dipbo, and M. Palmisano). Bureau of Labor Statistics, Washington, D.C., 129-138.
- MURPHY, L.R., and COWAN, C.D. (1976). Effects of bounding on telescoping in the National Crime Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 633-638.
- NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- NETER, J., and WAKSBERG, J. (1965). Response errors in collection of experimental data by household interviews: An experimental study. Technical Report No. 11, Bureau of the Census, Washington, D.C.
- SILBERSTEIN, A.R. (1988). Selected first-interview effects in the Consumer Expenditure Interview Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 485-490.
- SILBERSTEIN, A.R., and JACOBS, C.A. (1989). Symptoms of repeated interview effects in the Consumer Expenditure Interview Survey. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: Wiley, 289-303.
- SILBERSTEIN, A.R. (1989). Recall effects in the U.S. Consumer Expenditure Interview Survey. *Journal of Official Statistics*, 2, 125-142.
- SUDMAN, S., and BRADBURN N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.
- U.S. BUREAU OF LABOR STATISTICS (1986). *Consumer Expenditure Survey: Interview Survey, 1984*. Washington, D.C., Bulletin No. 2267.

Symmetry in Flows Among Reported Victimization Classifications with Nonresponse

ELIZABETH A. STASNY¹

ABSTRACT

The United States' National Crime Survey is a large-scale, household survey used to provide estimates of victimizations. The National Crime Survey uses a rotating panel design under which sampled housing units are maintained in the sample for three-and-one-half years with residents of the housing units being interviewed every six months. Nonresponse is a serious problem in longitudinal data from the National Crime Survey since as few as 25% of all individuals interviewed for the survey are respondents over an entire three-and-one-half-year period. In addition, the nonresponse typically does not occur at random with respect to victimization status. This paper presents models for gross flows among two types of victimization reporting classifications: number of victimizations and seriousness of victimization. The models allow for random or nonrandom nonresponse mechanisms, and allow the probabilities underlying the gross flows to be either unconstrained or symmetric. The models are fit, using maximum likelihood estimation, to the data from the National Crime Survey.

KEY WORDS: Categorical data; Ignorable nonresponse; Longitudinal survey; National Crime Survey; Nonignorable nonresponse.

1. INTRODUCTION

The United States' National Crime Survey (NCS) is a large-scale, household survey conducted by the U.S. Bureau of the Census for the Bureau of Justice Statistics. Data from the NCS is used to produce quarterly estimates of victimization rates and yearly estimates of the prevalence of crime. The survey uses a rotating panel of housing units (HU's) under which individuals living in sampled HU's are interviewed up to seven times at six-month intervals.

Individuals interviewed for the NCS are asked about crimes committed against them or against their property in the previous six months. In this work, we begin to explore the victimization status reported by households (HH's) within sampled HU's from one interview to the next. Victimization status for a HH will be considered in two ways: by the number of crimes reported (zero, one, and two or more) and by the type of crime reported (no crime, property crime, and personal contact crime).

Since responses are not available from one NCS interview period to the next for all HH's, we must decide how to handle missing observations. The nonresponse problem is a serious problem in the longitudinal data available from the NCS. For example, Fienberg (1980) noted that complete, three-and-one-half-year records of NCS interviews are available for as few as 25% of all individuals interviewed. In addition, the nonresponse typically does not occur at random with respect to victimization status (see, for example, Saphire (1984)).

This work extends the models developed by Stasny (1986) for nonrandom nonresponse in estimating gross flows. In particular, the models presented here allow for symmetry in the matrix of flows among victimization classifications as well as allowing for completely random nonresponse, ignorable nonrandom nonresponse, or nonignorable nonresponse.

¹ Elizabeth A. Stasny, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.

Section 2 of this paper provides a brief description of the NCS and the longitudinal data from the survey. Section 3 gives a general form of the models for symmetry in gross flow matrices with missing data and presents iterative procedures for obtaining maximum likelihood estimators (MLE's) for the parameters of the models. Section 4 describes the fits of the models to data from the NCS. Section 5 presents conclusions and suggests areas for future research.

2. THE NATIONAL CRIME SURVEY AND DATA

2.1 Survey Design

The NCS is a stratified, multi-stage, cluster sample of HU's. The survey was begun in July 1972 by the Law Enforcement Assistance Administration but has been administered by the Bureau of Justice Statistics since December 1979. The target population for the NCS is the civilian, non-institutionalized population of persons aged 12 and over living in housing units. The survey provides information on personal and household crimes committed against the individuals in sampled HU's. The following crimes and attempted crimes are covered by the NCS: assault, auto or motor vehicle theft, burglary, larceny, rape, and robbery. Crimes not covered by the survey include kidnapping, murder, shoplifting, and crimes that occur at places of business.

The NCS uses a rotating panel design under which a sampled HU is maintained in the sample for three and one-half years with interviews conducted at six-month intervals for a total of seven possible interviews. The initial interview at each HU, however, serves as a bounding interview and is not used for the purpose of estimation. Although there is a six-month interval between interviews at any one HU, NCS interviews are conducted in every month of the year; in order to make efficient use of trained interviewers, one-sixth of the HU's in the sample are scheduled for interviews each month. Since the sampling unit for the NCS is the HU, no attempt is made to follow individuals who move away from the HU during the three-and-one-half-year period. Rather, new individuals entering the HU are included in the survey. Each different group of individuals who live in a HU during its time in the NCS sample is considered a separate HH.

NCS interviews are conducted for all individuals 12 years of age or older who live in the sampled HU at the time of the interview. During the interview, individuals are asked about crimes committed against them or against the household in the previous six months. A single HH respondent is asked a series of six screening questions to elicit information on crimes committed against the HH (burglary, larceny, and motor vehicle theft). Then an eleven-question screener is used to elicit information from each individual in the HH concerning personal crimes committed against that individual (assault, rape, and robbery). An incident report is completed for each crime mentioned in response to the screening questions.

Additional information on the design and history of the NCS is provided, for example, by the U.S. Department of Justice and Bureau of Justice Statistics (1981), Saphire (1984), Dodge and Skogan (1987), and Montagliani (1987). A new sample design for the NCS has been used since January 1986. Taylor (1987) describes the redesign of the NCS and research associated with the redesign effort. The data used in this work, however, were collected under the original NCS design.

2.2 The Longitudinal Data

The data used in this work are from a large, longitudinal data set which includes all the regular NCS interview information collected from January 1975 to June 1979 except for the HU's that rotated into the sample in 1979. To make it easier to handle the data, this research uses only a subset of the data. The subset was created by taking a random start at the record

for the eighth HU in the full data set and then every fifteenth record after that. The resulting data set contains NCS records for 12,432 HU's. Because the HU's on the original longitudinal file are ordered in such a way that units from the same cluster appear together, the 1-in-15 systematic sample should not include two or more HU's from a single cluster. Thus, this research does not consider the problem of correlations among HU's within clusters.

2.3 Flows Among Victimization Classifications

The hierarchical, longitudinal data were used to create summary matrices for the years 1975, 1976, 1977, and 1978 showing flows among reported victimization classifications from each HH's first interview in a year to the HH's second interview for the year. Note that, since NCS interviews are conducted every month of the year, the first interview may occur at any time from January through June and the second interview may occur in July through December. Depending on the month of the interview, the victimizations reported in the first interview are those that occurred between the previous July and May while those reported in the second interview occurred between January and November. Thus, the analysis here explores only the reporting of crimes from one interview to the next. It cannot, for example, address issues of change in victimization reporting at various times of the year except in a very general sense.

It should be noted that during the time when the data were collected, a reference-period experiment was conducted using a sample of NCS HU's. Since individuals in HU's included in the experiment were asked to report victimizations for reference periods other than the usual six-month period, those HU's were not used in this analysis.

For the analyses here, each HH interviewed at least once during a given year was classified according to its reporting and victimization status at the two interview times. A victimization may have been reported by any member of the HH and may be against an individual or against the HH. Two sets of matrices showing victimization classifications are used in the analyses of Section 4. The matrices are given in Appendix I.

The first set of matrices show cross-classifications of HH's by the number of victimizations reported in the first and second interviews for each year. The classifications are: crime free (no victimizations reported), single crime (one victimization reported), multiple crime (two or more victimizations reported), and missing (HH did not respond or rotated out of the sample). The second set of matrices show cross-classifications of HH's by the type of victimization reported. The classifications are: crime free, property crime (burglary, larceny, and motor vehicle theft), contact crime (rape, assault, robbery, purse snatching, and pocket picking), and missing. These type-of-crime groupings are the same as those used in the NCS. In cases where multiple crimes were reported by a single HH, the classification used is for the most serious crime reported (contact crimes are taken to be more serious than property crimes).

Notice the large amount of nonresponse in the observed matrices shown in Appendix I. Only about 50% of the HH's who responded in at least one of the two interviews responded at both interview periods. The models presented in the following section, will allow us to handle this nonresponse while exploring the structure of the underlying matrix of probabilities of flows among the victimization classifications.

3. THE MODELS

This section presents a general form of the models that will be used to explore gross flows among victimization classifications in the NCS data. The form of the models follows that proposed by Chen and Fienberg (1974) for contingency tables with completely and partially classified data. The models for nonresponse are those developed by Stasny (1986) as well as

a model for random nonresponse. The model for symmetry in the flows, however, does not appear in the previous work. The models are presented in a general form because they are applicable to problems other than estimating gross flows among victimization classifications using NCS data.

3.1 Model for the Observed Data

Consider observation units that respond to a survey in at least one of two interview periods. Suppose that, when a unit responds to the survey, that unit is classified into one of K classifications. If a unit does not respond to the survey, that unit is classified as missing. Then the interview-to-interview flow data may be represented as in Table 1.

Table 1
Summary of Observed Data

		Time 2				Missing
		1	2	...	K	
Time 1	1	x_{11}	x_{12}	...	x_{1K}	x_{1M}
	2	x_{21}	x_{22}	...	x_{2K}	x_{2M}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	K	x_{K1}	x_{K2}	...	x_{KK}	x_{KM}
Missing		x_{M1}	x_{M2}	...	x_{MK}	

where x_{ij} = number of units with survey or missing status i at time 1 and j at time 2.

We suppose that each unit would fall into one of the cells of the $K \times K$ matrix of survey classifications if it were observed at both interview times. Let p_{ij} be the probability that a unit has status i at time 1 and status j at time 2, where i and j take on the values 1, 2, ..., K . Each unit in the (i,j) cell of the matrix of survey classifications has a chance of being missing at one of the two survey times. Let λ_{tij} be the probability that a unit in the (i,j) cell loses its classification at time t and, hence, is classified as missing at that time. Then the probabilities underlying the observed data are as shown in Table 2.

Table 2
Probabilities Underlying Observed Data

		Time 2				Missing
		1	2	...	K	
Time 1	1	$\{(1 - \lambda_{1ij} - \lambda_{2ij})p_{ij}\}$				$\left\{ \sum_{j=1}^K p_{ij} \lambda_{2ij} \right\}$
	2					
	\vdots					
	K					
Missing		$\left\{ \sum_{i=1}^K p_{ij} \lambda_{1ij} \right\}$				

Assuming that the p_{ij} are probabilities from a multinomial distribution, the likelihood function for the observed data is proportional to

$$\begin{aligned} &\left\{ \prod_{i=1}^K \prod_{j=1}^K [p_{ij} (1 - \lambda_{1ij} - \lambda_{2ij})]^{x_{ij}} \right\} \\ &\times \left\{ \prod_{i=1}^K \left[\sum_{j=1}^K p_{ij} \lambda_{2ij} \right]^{x_{iM}} \right\} \\ &\times \left\{ \prod_{j=1}^K \left[\sum_{i=1}^K p_{ij} \lambda_{1ij} \right]^{x_{Mj}} \right\}. \end{aligned}$$

There are $3K^2 + 2K - 1$ free parameters defined above and only $K^2 + 2K$ observed cells of data with a single constraint on the total sample size. Thus there are too many parameters to estimate using the observed data and we must reduce the number of parameters in the model. In the following we reduce the number of parameters to be estimated by considering two models for the p_{ij} -parameters and six models for the λ_{tij} -parameters.

3.2 Models for the p and λ Probabilities

We consider two models for the p_{ij} 's, the probabilities of flows among survey classifications: the unconstrained model and the model of symmetric flows. Under the model of unconstrained flow probabilities, there is a different probability, p_{ij} , for every (i,j) cell of the flow matrix. Under the model of symmetric flows, we have $p_{ij} = p_{ji}$ for $i \neq j$ so that the probability that a unit has survey classification i at time 1 and j at time 2 is the same as the probability that a unit has survey classification j at time 1 and i at time 2. Note that symmetry in the cell probabilities of the flow matrix implies equality of row and column marginal totals. Thus the model of symmetry in flow probabilities implies a certain stability in the population since the expected number of units with a particular survey classification at time 1 is the same as the number with that classification at time 2.

As defined above, the λ_{tij} 's, the probabilities that units with survey classifications i at time 1 and j at time 2 are missing at time t , depend on the time at which the nonresponse occurs and on the survey classifications at both times 1 and 2. We consider six simpler models for these probabilities. These models, along with the associated degrees of freedom under both models for the p_{ij} , are given below:

	d.f. unconstrained p_{ij}	d.f. symmetric p_{ij}
Model R: $\lambda_{tij} = \lambda$,	$2K - 1$	$(K^2 + 3K - 2)/2$
Model A: $\lambda_{1ij} = \lambda_{1j}$, $\lambda_{2ij} = \lambda_{2i}$,	0	$(K^2 - K)/2$
Model B: $\lambda_{tij} = \lambda_t$,	$2K - 2$	$(K^2 + 3K - 4)/2$
Model C: $\lambda_{1ij} = \lambda_j$, $\lambda_{2ij} = \lambda_i$,	K	$(K^2 + K)/2$
Model D: $\lambda_{1ij} = \lambda_{1i}$, $\lambda_{2ij} = \lambda_{2j}$,	0	$(K^2 - K)/2$
Model E: $\lambda_{1ij} = \lambda_i$, $\lambda_{2ij} = \lambda_j$,	K	$(K^2 + K)/2$

Model R is the model of random nonresponse. Under Model R, there is a single probability of nonresponse for all units at both times regardless of survey classification. Under Model A, the probability that a unit is missing at time t depends on both the time and the survey classification at the time when the unit responds. Note that if Model A is used for the λ -parameters and the unconstrained model is used for the p_{ij} , then the model is a saturated model which will fit the data exactly. Under Model B, the probability that a unit is missing at time t depends only on the time. Under Model C, the probability that a unit is missing at time t depends only on the unit's survey classification at the time when the unit responds. Under Model D, the probability that a unit is missing at time t depends on both the time and the survey classification at the time when the unit is missing. If Model D is used for the λ -parameters and the unconstrained model is used for the p_{ij} , then the model is a saturated model which will fit the data exactly. Under Model E, the probability that a unit is missing at time t depends only on the unit's survey classification at the time when the unit is missing.

Under Model R, nonresponse is said to be completely at random. Under Models A, B, and C, nonresponse is said to be ignorable nonresponse in that the nonresponse mechanism depends only on the observed data. Nonresponse under Models D and E is nonignorable nonresponse since the nonresponse mechanism depends on the missing data. (See Little and Rubin (1987) for more information on the types of nonresponse.)

In the following two subsections, we describe procedures for fitting the models presented above. The fits of the models can be assessed using either the Pearson X^2 statistic or G^2 , the likelihood ratio statistic. Both statistics have asymptotic χ^2 distributions, with degrees of freedom as shown above, given that the model is correct. In the following we use the notation "Model R-U" to denote the pairing of Model R for the λ -parameters and the unconstrained model for the p_{ij} . "Model R-S" will denote the pairing of Model R for the λ -parameters and the symmetric model for the p_{ij} . Similar notation will be used to denote the pairings of Models A, B, C, D, and E for the λ -parameters with one of the two models for the p_{ij} .

3.3 Estimation of the p and λ Parameters Under Models R, A, B, and C

The likelihood functions for the eight models created using one of the two models for the p_{ij} and Model R, A, B, or C for the λ_{tij} factor into two pieces: one piece a function of the p -parameters alone and one a function of the λ -parameters alone. Thus, the MLE's may be found separately for the two sets of parameters. In addition, the p -parameter estimates do not depend on which of these four models is used for the λ -parameters, and the λ -parameter estimates do not depend on which of the two models is used for the p -parameters.

An iterative procedure for obtaining MLE's for the p -parameters under the unconstrained model paired with Model R, A, B, or C for the λ -parameters is given in Chen and Fienberg (1974). The equations for this procedure are provided in Appendix II.

Under the symmetric model for the p -parameters paired with Model R, A, B, or C for the λ -parameters, the factor of the likelihood equation involving only the p_{ij} 's is as follows:

$$\left\{ \prod_{i=1}^k p_{ii}^{x_{ii}} \right\} \times \left\{ \prod_{i=1}^k \prod_{j=i+1}^k p_{ij}^{x_{ij}} \right\} \times \left\{ \prod_{i=2}^k \prod_{j=1}^{i-1} p_{ji}^{x_{ji}} \right\} \\ \times \left\{ \prod_{i=1}^k p_{i \cdot}^{x_{iM}} \right\} \times \left\{ \prod_{j=1}^k p_{j \cdot}^{x_{jM}} \right\}, \quad (1)$$

where a dot in a subscript indicates summation over that subscript. Equation (1) is maximized subject to the constraint that the sum of the p_{ij} 's is one. In general, an iterative procedure is required to obtain the MLE's. Let $x_{..} = \sum_{i=1}^K \sum_{j=1}^K x_{ij}$ be the total number of units observed at both times and let $n = x_{..} + x_{.M} + x_{M.}$ be the total number of units observed in at least one of the two interview times. Then the iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Symmetric p_{ij} Under Models R, A, B, and C

1. $p_{ii}^{(0)} = x_{ii}/x_{..}$
 $p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..} \quad \text{for } i \neq j.$
2. $p_{ii}^{(v+1)} = [x_{ii} + (x_{iM} + x_{Mi})p_{ii}^{(v)}/p_{i.}^{(v)}]/n$
 $p_{ij}^{(v+1)} = [(x_{ij} + x_{ji}) + (x_{iM} + x_{Mi})p_{ij}^{(v)}/p_{i.}^{(v)} + (x_{jM} + x_{Mj})p_{ij}^{(v)}/p_{j.}^{(v)}]/2n \quad \text{for } i \neq j.$

Step 2 is repeated for $v = 0, 1, 2, \dots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other positive values satisfying the constraint that the p_{ij} 's sum to one may be used.

An iterative procedure for obtaining MLE's for the λ -parameters under Model A and the closed-form estimator for the λ -parameters under Model B are given in Chen and Fienberg (1974). An iterative procedure for obtaining MLE's for the λ -parameters under Model C is given in Stasny (1986). The equations for these procedures are provided in Appendix II.

Under Model R for the λ -parameters, the factor of the likelihood equation involving only λ is as follows:

$$\left\{ \prod_{i=1}^K \prod_{j=1}^K (1 - 2\lambda)^{x_{ij}} \right\} \times \left\{ \prod_{i=1}^K \lambda^{x_{iM}} \right\} \times \left\{ \prod_{j=1}^K \lambda^{x_{Mj}} \right\}.$$

The closed-form MLE for λ is

$$\hat{\lambda} = (x_{.M} + x_{M.})/2n.$$

3.4 Estimation of the p and λ Parameters Under Model D

The likelihood functions for the observed data under either Model D-U or Model D-S cannot be factored and all parameter estimates must be obtained simultaneously. An iterative procedure for obtaining MLE's under Model D-U is given in Stasny (1988). The equations for this procedure are provided in Appendix II. Under Model D-S, the likelihood function for the observed data is as follows:

$$\begin{aligned} & \left\{ \prod_{i=1}^K p_{ii}^{x_{ii}} \right\} \times \left\{ \prod_{i=1}^K \prod_{j=i+1}^K p_{ij}^{x_{ij}} \right\} \times \left\{ \prod_{i=2}^K \prod_{j=i}^{i-1} p_{ji}^{x_{ji}} \right\} \times \left\{ \prod_{i=1}^K \prod_{j=1}^K [(1 - \lambda_{1i} - \lambda_{2j})]^{x_{ij}} \right\} \\ & \times \left\{ \prod_{i=1}^K \left[\sum_{j=1}^K p_{ij} \lambda_{2j} \right]^{x_{iM}} \right\} \times \left\{ \prod_{j=1}^K \left[\sum_{i=1}^K p_{ji} \lambda_{1i} \right]^{x_{Mj}} \right\}. \end{aligned} \quad (2)$$

Equation (2) is maximized subject to the constraint that the sum of the p_{ij} 's is one. In general, an iterative procedure is required in order to obtain the MLE's. The iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Parameters Under Model D-S

$$1. p_{ii}^{(0)} = x_{ii}/x_{..}$$

$$p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..} \quad \text{for } i \neq j$$

$$\lambda_{1i}^{(0)} = x_{M\cdot}/n$$

$$\lambda_{2j}^{(0)} = x_{\cdot M}/n.$$

$$2. p_{ii}^{(\nu+1)} = n^{-1} \left\{ x_{ii} + x_{iM} \left[p_{ii}^{(\nu)} \lambda_{2i}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{2h}^{(\nu)} \right] + x_{Mi} \left[p_{ii}^{(\nu)} \lambda_{1i}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{1h}^{(\nu)} \right] \right\}$$

$$\begin{aligned} p_{ij}^{(\nu+1)} = (2n)^{-1} & \left\{ x_{ij} + x_{ji} + x_{iM} \left[p_{ij}^{(\nu)} \lambda_{2j}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{2h}^{(\nu)} \right] \right. \\ & + x_{jM} \left[p_{ij}^{(\nu)} \lambda_{2i}^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_{2h}^{(\nu)} \right] + x_{Mi} \left[p_{ij}^{(\nu)} \lambda_{1j}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{1h}^{(\nu)} \right] \\ & \left. + x_{Mj} \left[p_{ij}^{(\nu)} \lambda_{1i}^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_{1h}^{(\nu)} \right] \right\} \quad \text{for } i \neq j \end{aligned}$$

$$\lambda_{1i}^{(\nu+1)} = \sum_{j=1}^K \left[x_{Mj} p_{ij}^{(\nu)} \lambda_{1i}^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_{1h}^{(\nu)} \right] / \sum_{j=1}^K [x_{ij} / (1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)})]$$

$$\lambda_{2j}^{(\nu+1)} = \sum_{i=1}^K \left[x_{iM} p_{ij}^{(\nu)} \lambda_{2j}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{2h}^{(\nu)} \right] / \sum_{i=1}^K [x_{ij} / (1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)})].$$

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other values between zero and one satisfying the constraint that the p_{ij} 's sum to one may be used.

3.5 Estimation of the p and λ Parameters Under Model E

The likelihood functions for the observed data under either Model E-U or Model E-S cannot be factored and all parameter estimates must be obtained simultaneously. An iterative procedure for obtaining MLE's under Model E-U is given in Stasny (1988). The equations for this procedure are provided in Appendix II. Under Model E-S, the likelihood function for the observed data is as follows:

$$\left\{ \prod_{i=1}^K p_{ii}^{x_{ii}} \right\} \times \left\{ \prod_{i=1}^K \prod_{j=i+1}^K p_{ij}^{x_{ij}} \right\} \times \left\{ \prod_{i=2}^K \prod_{j=1}^{i-1} p_{ji}^{x_{ij}} \right\} \times \left\{ \prod_{i=1}^K \prod_{j=1}^K [(1 - \lambda_i - \lambda_j)]^{x_{ij}} \right\} \\ \times \left\{ \prod_{i=1}^K \left[\sum_{j=1}^K p_{ij} \lambda_j \right]^{x_{iM}} \right\} \times \left\{ \prod_{j=1}^K \left[\sum_{i=1}^K p_{ji} \lambda_i \right]^{x_{Mj}} \right\}. \quad (3)$$

Equation (3) is maximized subject to the constraint that the sum of the p_{ij} 's is one. In general, an iterative procedure is required in order to obtain the MLE's. The iterative procedure used in the data analysis reported in Section 4 is as follows:

Iterative Procedure for Estimating Parameters Under Model E-S

$$1. p_{ii}^{(0)} = x_{ii}/x_{..}$$

$$p_{ij}^{(0)} = (x_{ij} + x_{ji})/2x_{..} \quad \text{for } i \neq j$$

$$\lambda_i^{(0)} = (x_{M.} + x_{.M})/2n.$$

$$2. p_{ii}^{(\nu+1)} = n^{-1} \left\{ x_{ii} + (x_{iM} + x_{Mi}) \left[p_{ii}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_h^{(\nu)} \right] \right\}$$

$$p_{ij}^{(\nu+1)} = (2n)^{-1} \left\{ x_{ij} + x_{ji} + (x_{iM} + x_{Mi}) \left[p_{ij}^{(\nu)} \lambda_j^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_h^{(\nu)} \right] \right. \\ \left. + (x_{jM} + x_{Mj}) \left[p_{ij}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_h^{(\nu)} \right] \right\} \quad \text{for } i \neq j$$

$$\lambda_i^{(\nu+1)} = \sum_{j=1}^K \left[(x_{jM} + x_{Mj}) p_{ji}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_h^{(\nu)} \right] \\ \left/ \sum_{j=1}^K [(x_{ij} + x_{ji}) / (1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)})] \right.$$

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the parameter estimates converge to the desired degree of accuracy. The initial estimates given in step 1 are merely suggested estimates. Other values between zero and one satisfying the constraint that the p_{ij} 's sum to one may be used.

4. FITS OF THE MODELS TO NCS DATA

The models described in Section 3 were fit to the NCS data described in Section 2. Recall that the NCS data for each of the years from 1975 to 1978 is summarized both by number of crimes reported in each of the two interviews during the year and by the type of crime reported. Since three survey classifications are used, we have $K = 3$. Standard errors of the parameter estimates were obtained using the observed information matrix.

Table 3a
Estimates of p_{ij} for Flows Among Number-of-Crime Classifications
Under Models R, A, B, and C

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Single Crime	Multiple Crime	Crime Free	Single Crime	Multiple Crime
1975							
First Interview	Crime Free	.666 (.0075)	.098 (.0050)	.029 (.0031)	.666 (.0075)	.102 (.0035)	.032 (.0022)
	Single Crime	.106 (.0051)	.029 (.0031)	.014 (.0023)	.102 (.0035)	.029 (.0031)	.012 (.0015)
	Multiple Crime	.036 (.0032)	.011 (.0021)	.012 (.0021)	.032 (.0022)	.012 (.0015)	.012 (.0021)
1976							
First Interview	Crime Free	.669 (.0076)	.101 (.0052)	.029 (.0033)	.669 (.0076)	.099 (.0036)	.030 (.0022)
	Single Crime	.098 (.0051)	.034 (.0034)	.014 (.0025)	.099 (.0036)	.034 (.0034)	.014 (.0017)
	Multiple Crime	.031 (.0030)	.014 (.0023)	.011 (.0022)	.030 (.0022)	.014 (.0017)	.010 (.0022)
1977							
First Interview	Crime Free	.670 (.0079)	.115 (.0058)	.032 (.0034)	.671 (.0079)	.103 (.0037)	.030 (.0023)
	Single Crime	.092 (.0051)	.026 (.0032)	.016 (.0026)	.103 (.0037)	.026 (.0032)	.016 (.0018)
	Multiple Crime	.028 (.0030)	.016 (.0026)	.006 (.0017)	.030 (.0023)	.016 (.0018)	.006 (.0017)
1978							
First Interview	Crime Free	.671 (.0087)	.097 (.0062)	.027 (.0035)	.671 (.0087)	.105 (.0043)	.027 (.0025)
	Single Crime	.111 (.0061)	.032 (.0040)	.009 (.0022)	.105 (.0043)	.032 (.0040)	.010 (.0017)
	Multiple Crime	.027 (.0034)	.013 (.0027)	.013 (.0026)	.027 (.0025)	.010 (.0017)	.013 (.0026)

Note: Estimated standard errors are given in parentheses.

4.1 Estimates of the p -Parameters Under Models R, A, B, and C

Recall that the p -parameter estimates do not depend on the nonresponse mechanism under Models R, A, B, and C. For the iterative procedures used to estimate the p_{ij} under both the unconstrained and symmetric models, the criterion used for stopping the iteration was that the expected counts in the (i, j) cell of the flow matrix, $n\hat{p}_{ij}$, differed by no more than 0.5 from one step of the iterative procedure to the next. In all cases, convergence occurred rapidly, taking at most six steps. The estimates of the p_{ij} when HH's are classified by numbers of crimes reported are given in Table 3a for both the unconstrained and symmetric models. The estimates of the p_{ij} when HH's are classified by types of crimes reported are given in Table 4a for both the unconstrained and symmetric models.

Table 3b
Estimates of p_{ij} for Flows Among Number-of-Crime Classifications
Under Models D-S

		Symmetric Model		
		Second Interview		
		Crime Free	Single Crime	Multiple Crime
1975				
First Interview	Crime Free	.638 (.0104)	.106 (.0047)	.035 (.0029)
	Single Crime	.106 (.0047)	.033 (.0039)	.015 (.0019)
	Multiple Crime	.035 (.0029)	.015 (.0019)	.016 (.0027)
1976				
First Interview	Crime Free	.645 (.0100)	.100 (.0045)	.034 (.0029)
	Single Crime	.100 (.0045)	.037 (.0041)	.017 (.0021)
	Multiple Crime	.034 (.0029)	.017 (.0021)	.015 (.0029)
1977				
First Interview	Crime Free	.642 (.0109)	.106 (.0054)	.033 (.0032)
	Single Crime	.106 (.0054)	.031 (.0043)	.021 (.0023)
	Multiple Crime	.033 (.0032)	.021 (.0023)	.009 (.0025)
1978				
First Interview	Crime Free	.636 (.0118)	.114 (.0056)	.028 (.0029)
	Single Crime	.114 (.0056)	.040 (.0051)	.013 (.0021)
	Multiple Crime	.028 (.0029)	.013 (.0021)	.015 (.0030)

Note: Estimated standard errors are given in parentheses.

Notice in both Tables 3a and 4a that the flow matrices of estimated probabilities under the unconstrained model for the p_{ij} appear to be fairly symmetric so that the model of symmetry in the flows is suggested as a reasonable model to consider. Also notice that the estimates of the p_{ij} do not appear to change much over the four years. The fits of these two models for the p_{ij} will be considered for each of the four models for nonresponse in Subsection 4.4 below.

Table 3c
Estimates of p_{ij} for Flows Among Number-of-Crime Classifications
Under Models E-U and E-S

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Single Crime	Multiple Crime	Crime Free	Single Crime	Multiple Crime
1975							
First Interview	Crime Free	.639 (.0104)	.102 (.0061)	.031 (.0037)	.639 (.0104)	.106 (.0047)	.035 (.0028)
	Single Crime	.110 (.0061)	.033 (.0039)	.016 (.0026)	.106 (.0047)	.033 (.0039)	.015 (.0019)
	Multiple Crime	.039 (.0039)	.014 (.0025)	.016 (.0027)	.035 (.0028)	.015 (.0019)	.016 (.0027)
1976							
First Interview	Crime Free	.645 (.0100)	.103 (.0063)	.032 (.0041)	.645 (.0100)	.101 (.0045)	.033 (.0029)
	Single Crime	.098 (.0057)	.037 (.0041)	.017 (.0030)	.101 (.0045)	.037 (.0041)	.017 (.0021)
	Multiple Crime	.035 (.0037)	.017 (.0027)	.016 (.0029)	.033 (.0029)	.017 (.0021)	.016 (.0029)
1977							
First Interview	Crime Free	.636 (.0112)	.124 (.0083)	.037 (.0050)	.642 (.0110)	.106 (.0055)	.033 (.0033)
	Single Crime	.094 (.0060)	.031 (.0043)	.021 (.0031)	.106 (.0055)	.030 (.0043)	.020 (.0023)
	Multiple Crime	.029 (.0036)	.020 (.0031)	.008 (.0024)	.033 (.0033)	.020 (.0023)	.008 (.0025)
1978							
First Interview	Crime Free	.639 (.0118)	.106 (.0078)	.029 (.0042)	.637 (.0118)	.112 (.0055)	.028 (.0029)
	Single Crime	.117 (.0070)	.041 (.0051)	.011 (.0026)	.112 (.0055)	.041 (.0051)	.013 (.0021)
	Multiple Crime	.027 (.0037)	.016 (.0032)	.015 (.0030)	.028 (.0029)	.013 (.0021)	.015 (.0030)

Note: Estimated standard errors are given in parentheses.

Table 4a
Estimates of p_{ij} for Flows Among Type-of-Crime Classifications
Under Models R, A, B, and C

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Property Crime	Contact Crime	Crime Free	Property Crime	Contact Crime
1975							
First Interview	Crime Free	.666 (.0075)	.105 (.0053)	.022 (.0026)	.666 (.0075)	.111 (.0037)	.024 (.0018)
	Property Crime	.118 (.0054)	.044 (.0038)	.010 (.0019)	.111 (.0037)	.044 (.0038)	.008 (.0013)
	Contact Crime	.025 (.0026)	.007 (.0016)	.004 (.0012)	.024 (.0018)	.008 (.0013)	.004 (.0012)
1976							
First Interview	Crime Free	.669 (.0076)	.108 (.0055)	.023 (.0028)	.669 (.0021)	.108 (.0011)	.022 (.0010)
	Property Crime	.108 (.0053)	.047 (.0040)	.010 (.0021)	.108 (.0011)	.047 (.0019)	.011 (.0009)
	Contact Crime	.021 (.0025)	.012 (.0021)	.002 (.0011)	.022 (.0010)	.011 (.0009)	.002 (.0012)
1977							
First Interview	Crime Free	.670 (.0079)	.128 (.0061)	.019 (.0026)	.671 (.0078)	.115 (.0039)	.018 (.0018)
	Property Crime	.103 (.0053)	.041 (.0039)	.008 (.0018)	.115 (.0039)	.041 (.0040)	.008 (.0014)
	Contact Crime	.016 (.0025)	.008 (.0021)	.006 (.0018)	.018 (.0018)	.008 (.0014)	.006 (.0017)
1978							
First Interview	Crime Free	.671 (.0087)	.104 (.0064)	.019 (.0031)	.671 (.0088)	.112 (.0044)	.019 (.0021)
	Property Crime	.119 (.0063)	.040 (.0044)	.010 (.0024)	.112 (.0044)	.040 (.0044)	.010 (.0017)
	Contact Crime	.019 (.0029)	.011 (.0025)	.006 (.0020)	.019 (.0021)	.010 (.0017)	.006 (.0020)

Note: Estimated standard errors are given in parentheses.

Table 4b
Estimates of p_{ij} for Flows Among Type-of-Crime Classifications
Under Models D-S

		Symmetric Model		
		Second Interview		
		Crime Free	Property Crime	Contact Crime
1975				
First Interview	Crime Free	.635 (.0101)	.118 (.0046)	.026 (.0026)
	Property Crime	.118 (.0046)	.052 (.0046)	.011 (.0016)
	Contact Crime	.026 (.0026)	.011 (.0016)	.005 (.0016)
1976				
First Interview	Crime Free	.641 (.0098)	.110 (.0046)	.026 (.0028)
	Property Crime	.110 (.0046)	.052 (.0048)	.015 (.0021)
	Contact Crime	.026 (.0028)	.015 (.0021)	.004 (.0019)
1977				
First Interview	Crime Free	.642 (.0104)	.120 (.0052)	.019 (.0024)
	Property Crime	.120 (.0052)	.050 (.0049)	.011 (.0019)
	Contact Crime	.019 (.0024)	.011 (.0019)	.008 (.0022)
1978				
First Interview	Crime Free	.636 (.0117)	.121 (.0057)	.020 (.0025)
	Property Crime	.121 (.0057)	.049 (.0055)	.012 (.0021)
	Contact Crime	.020 (.0025)	.012 (.0021)	.008 (.0025)

Note: Estimated standard errors are given in parentheses.

Table 4c
Estimates of p_{ij} for Flows Among Type-of-Crime Classifications
Under Models E-U and E-S

		Unconstrained Model			Symmetric Model		
		Second Interview					
		Crime Free	Property Crime	Contact Crime	Crime Free	Property Crime	Contact Crime
1975							
First Interview	Crime Free	.636 (.0100)	.111 (.0062)	.024 (.0034)	.636 (.0101)	.117 (.0046)	.026 (.0026)
	Property Crime	.124 (.0063)	.053 (.0047)	.012 (.0023)	.117 (.0046)	.052 (.0047)	.011 (.0016)
	Contact Crime	.027 (.0033)	.009 (.0020)	.005 (.0016)	.026 (.0026)	.011 (.0016)	.005 (.0016)
1976							
First Interview	Crime Free	.641 (.0098)	.110 (.0065)	.028 (.0041)	.641 (.0098)	.110 (.0046)	.026 (.0028)
	Property Crime	.110 (.0059)	.051 (.0048)	.014 (.0028)	.110 (.0046)	.052 (.0048)	.015 (.0021)
	Contact Crime	.024 (.0033)	.016 (.0028)	.005 (.0019)	.026 (.0028)	.015 (.0021)	.005 (.0019)
1977							
First Interview	Crime Free	.636 (.0108)	.138 (.0076)	.023 (.0035)	.641 (.0105)	.121 (.0051)	.019 (.0024)
	Property Crime	.107 (.0060)	.050 (.0048)	.010 (.0022)	.121 (.0051)	.049 (.0048)	.011 (.0018)
	Contact Crime	.015 (.0028)	.011 (.0027)	.009 (.0023)	.019 (.0024)	.011 (.0018)	.009 (.0022)
1978							
First Interview	Crime Free	.641 (.0117)	.111 (.0078)	.022 (.0040)	.640 (.0117)	.118 (.0056)	.021 (.0026)
	Property Crime	.124 (.0071)	.048 (.0055)	.012 (.0029)	.118 (.0056)	.048 (.0054)	.013 (.0021)
	Contact Crime	.020 (.0033)	.014 (.0031)	.009 (.0025)	.021 (.0026)	.013 (.0021)	.008 (.0025)

Note: Estimated standard errors are given in parentheses.

4.2 Estimates of the λ -Parameters Under Models R, A, B, and C

Recall that the λ -parameter estimates under Models R, A, B, and C are the same regardless of whether the unconstrained or symmetric model is used for the p -parameters. For the iterative procedures used to estimate the λ -parameters under Models A and C, the convergence criterion used was that estimates of the λ -parameters differed by no more than .0005 from one step to the next. Convergence took between 41 and 4150 steps when it occurred in fewer than 10,000 steps after using the initial parameter estimates suggested in Appendix II. The factors of the likelihood for the observed data involving only the λ -parameters were, in some cases, not well behaved. This is particularly true for the likelihoods for the 1978 data under both Models A and C. In such cases, a grid search was used to locate appropriate starting points for the iterative procedures. A rough grid search was also used in all cases to verify that, when the iterative procedure converged, it appeared to have converged to a global rather than a local maximum.

The estimates of the λ -parameters under both the number-of-crimes and type-of-crime classifications for Models R, A, B, and C are given in Tables 5, 6, 7, and 8 respectively.

Notice that under Models R and B the estimates of the λ -parameters are the same for both the number-of-crimes and type-of-crime classifications because the probability of being a nonrespondent under those two models does not depend on survey classification. Under Models A and C, the λ -parameter estimates corresponding to the crime-free classification are the same, within rounding error, for both the number-of-crimes and type-of-crime classifications since crime-free HH's are the same under both classifications. Also notice that, under Models A and C, the λ -parameter estimates, the estimated probabilities of being a nonrespondent, generally increase as the number of victimizations or the seriousness of the crime increases.

Table 5
Estimates of λ Under Model R

	Number-of-Crimes or Type-of-Crime Classification of Data
	$\hat{\lambda}$
1975	.224 (.0035)
1976	.232 (.0035)
1977	.237 (.0036)
1978	.250 (.0040)

Note: Estimated standard errors are given in parentheses.

Table 6
Estimates of λ_{1j} and λ_{2i} Under Model A

	Number-of-Crimes Classification of Data						Type-of-Crime Classification of Data					
	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$
1975	.208 (.0062)	.272 (.0159)	.327 (.0261)	.221 (.0064)	.234 (.0147)	.275 (.0242)	.208 (.0062)	.280 (.0151)	.322 (.0321)	.220 (.0064)	.246 (.0139)	.246 (.0303)
1976	.206* (.0063)	.261* (.0152)	.397* (.0268)	.236* (.0066)	.254* (.0153)	.267* (.0248)	.206 (.0063)	.278 (.0146)	.381 (.0327)	.235 (.0066)	.253 (.0144)	.285 (.0319)
1977	.192 (.0064)	.263 (.0152)	.309 (.0265)	.258 (.0070)	.281 (.0171)	.326 (.0285)	.192 (.0064)	.275 (.0144)	.267 (.0327)	.258 (.0069)	.269 (.0159)	.417 (.0369)
1978	.207* (.0072)	.316* (.0182)	.302* (.0308)	.269* (.0079)	.280* (.0176)	.321* (.0300)	.207* (.0072)	.305* (.0174)	.343* (.0364)	.269* (.0079)	.280* (.0166)	.334* (.0362)

Note: * Indicates cases in which the likelihood function is not well behaved.
Estimated standard errors are given in parentheses.

Table 7
Estimates of λ_1 and λ_2 Under Model B

	Number-of-Crimes or Type-of-Crime Classification of Data	
	$\hat{\lambda}_1$	$\hat{\lambda}_2$
1975	.223 (.0058)	.226 (.0058)
1976	.225 (.0059)	.240 (.0060)
1977	.209 (.0059)	.264 (.0064)
1978	.227 (.0067)	.273 (.0071)

Note: Estimated standard errors are given in parentheses.

Table 8
Estimates of λ_i Under Model C

	Number-of-Crimes Classification of Data			Type-of-Crime Classification of Data		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
1975	.214 (.0039)	.252 (.0118)	.300 (.0199)	.214 (.0039)	.262 (.0109)	.284 (.0262)
1976	.221 (.0040)	.257 (.0116)	.330 (.0210)	.221 (.0040)	.266 (.0109)	.333 (.0289)
1977	.225 (.0041)	.271 (.0126)	.317 (.0235)	.225* (.0041)	.273* (.0115)	.339* (.0286)
1978	.237* (.0046)	.297* (.0139)	.312* (.0236)	.237* (.0046)	.292* (.0130)	.339* (.0299)

Note: * Indicates cases in which the likelihood function is not well behaved.
Estimated standard errors are given in parentheses.

4.3 Parameter Estimates Under Models D and E

Models D and E are more difficult to fit than Models R, A, B, and C because all parameters under Models D and E must be estimated simultaneously. For all sets of the NCS data, the likelihood functions under Models D and E were not well behaved and grid searches over the possible values of the λ -parameters were required to locate suitable starting points for the iterative procedure. Since a grid search over the six λ -parameters under Model D was extremely time-consuming, parameter estimates were obtained under Model D-S but not under Model D-U. Estimates of the p -parameters under Model D-S are given in Table 3b for the number-of-crimes classification and in Table 4b for the type-of-crime classification. The λ -parameter estimates under Model D-S are given in Table 9 for both types of classifications. Estimates of the p -parameters under Models E-U and E-S are given in Table 3c for the number-of-crimes classification and in Table 4c for the type-of-crime classification. The λ -parameter estimates under Models E-U and E-S are given in Table 10 for both types of classifications.

Notice that under Models D and E the estimates of p_{11} , the probability of remaining in the crime-free classification, are somewhat smaller than the corresponding estimates under Models R, A, B, and C; the estimates of the remaining p -parameters under Models D and E are somewhat larger than the corresponding estimates under Models R, A, B, and C. Under both Models D and E, the λ -parameter estimates, the estimated probabilities of being a nonrespondent, generally increase as the number of victimizations or the seriousness of the crime increases. In the cases where the estimates decrease as the number of victimizations or the seriousness of the crime increases (in the 1978 data under Model D-S and in the 1978 number-of-crimes data under Model E-S), the decreases are small and within the estimated standard error of the estimates.

Table 9
Estimates of λ_{1i} and λ_{2j} Under Model D-S

	Number-of-Crimes Classification of Data						Type-of-Crime Classification of Data					
	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{13}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{22}$	$\hat{\lambda}_{23}$
1975	.210 (.0085)	.246 (.0303)	.319 (.0368)	.194 (.0085)	.321 (.0282)	.387 (.0362)	.208 (.0084)	.264 (.0249)	.319 (.0523)	.192 (.0085)	.339 (.0235)	.372 (.0507)
1976	.204 (.0083)	.276 (.0274)	.339 (.0344)	.217 (.0084)	.273 (.0291)	.444 (.0331)	.203 (.0083)	.280 (.0244)	.383 (.0443)	.215 (.0084)	.297 (.0255)	.453 (.0416)
1977	.175 (.0086)	.307 (.0301)	.380 (.0403)	.249 (.0089)	.298 (.0326)	.374 (.0439)	.175 (.0086)	.304 (.0243)	.438 (.0424)	.248 (.0089)	.315 (.0259)	.341 (.0491)
1978	.211 (.0094)	.278 (.0282)	.290 (.0433)	.236 (.0099)	.413 (.0261)	.384 (.0443)	.211 (.0094)	.276 (.0264)	.293 (.0563)	.236 (.0098)	.411 (.0246)	.391 (.0567)

Note: Estimated standard errors are given in parentheses.

Table 10
Estimates of λ_i Under Model E

	Number-of-Crimes Classification of Data			Type-of-Crime Classification of Data		
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
Unconstrained p_{ij}						
1975	.202 (.0060)	.285 (.0235)	.348 (.0262)	.201 (.0058)	.302 (.0180)	.336 (.0418)
1976	.211 (.0057)	.275 (.0226)	.387 (.0232)	.209 (.0056)	.286 (.0193)	.419 (.0327)
1977	.210 (.0063)	.315 (.0259)	.372 (.0351)	.209 (.0061)	.318 (.0183)	.394 (.0295)
1978	.224 (.0065)	.340 (.0208)	.342 (.0296)	.225 (.0065)	.326 (.0203)	.385 (.0333)
Symmetric p_{ij}						
1975	.202 (.0060)	.285 (.0235)	.351 (.0258)	.201 (.0059)	.301 (.0180)	.341 (.0408)
1976	.211 (.0057)	.274 (.0223)	.389 (.0229)	.209 (.0056)	.287 (.0191)	.418 (.0327)
1977	.213 (.0061)	.301 (.0267)	.376 (.0339)	.213 (.0060)	.309 (.0190)	.391 (.0302)
1978	.224 (.0065)	.343 (.0204)	.338 (.0298)	.225 (.0065)	.329 (.0199)	.379 (.0339)

Note: Estimated standard errors are given in parentheses.

4.4 Fits of the Models

Table 11 shows the X^2 and G^2 values and the associated degrees of freedom for all twelve models (including Model D-U which must fit the data exactly) and both types of survey classifications. Note that the models were fit as an illustration of the methods developed here and we have ignored the complex survey design. Although clusters are not a problem in our subsample of the NCS data, in a more complete analysis we would prefer to fit the models separately to data from different strata and then combine the strata estimates to obtain estimates for the entire population.

Clearly, neither Model R, the model of random nonresponse, nor Model B, under which the probability of nonresponse depends only on time, fits the data well for either the unconstrained or symmetric models for the p_{ij} .

Models C-U and C-S fit the 1975 data fairly well and give reasonable fits to the 1976 data. Since Model C-S fits the data reasonably well and is a more parsimonious model, we prefer it over Model C-U. Under Model C, the probability of nonresponse depends only on the victimization classification at the interview in which the HH responded, not on the time. Thus, Model C is the model of symmetry in the nonresponse probabilities for the two interview periods. When Model C is paired with the symmetric model for the p -parameters, we obtain symmetric expected cell counts for the observed flow data. Notice in the observed data shown in Appendix I, that in 1977 and 1978 there is much more nonresponse at the second interview time than at the first interview time. This difference in nonresponse rates is the reason for the lack of fit of Model C to the 1977 and 1978 data.

Table 11
Fits of the Models

Number-of-Crimes Classification of Data					Type-of-Crime Classification of Data				
Unconstrained p_{ij}		Symmetric p_{ij}			Unconstrained p_{ij}		Symmetric p_{ij}		
X^2	G^2	X^2	G^2		X^2	G^2	X^2	G^2	
Model R									
(d.f. = 5)		(d.f. = 8)			(d.f. = 5)		(d.f. = 8)		
1975	42.7	41.2	45.9	45.6	38.2	36.9	42.0	41.5	
1976	70.2	67.1	69.7	67.7	57.7	55.9	58.3	56.4	
1977	74.2	75.2	83.9	85.3	85.4	84.8	94.8	95.3	
1978	61.7	62.7	64.9	66.3	63.2	64.1	65.5	66.8	
Model A									
(d.f. = 0)		(d.f. = 3)			(d.f. = 0)		(d.f. = 3)		
1975	0.0	0.0	4.4	4.4	0.0	0.0	4.6	4.6	
1976	0.0	0.0	0.6	0.6	0.0	0.0	0.5	0.5	
1977	0.0	0.0	10.1	10.1	0.0	0.0	10.5	10.5	
1978	0.0	0.0	3.7	3.7	0.0	0.0	2.7	2.7	
Model B									
(d.f. = 4)		(d.f. = 7)			(d.f. = 4)		(d.f. = 7)		
1975	42.7	41.1	45.9	45.5	38.2	36.9	42.0	41.5	
1976	69.1	64.5	68.5	65.1	56.2	53.3	56.9	53.8	
1977	47.1	45.4	58.7	55.5	57.0	54.9	68.4	65.4	
1978	47.6	46.0	50.1	49.6	49.1	47.4	50.7	50.1	
Model C									
(d.f. = 3)		(d.f. = 6)			(d.f. = 3)		(d.f. = 6)		
1975	6.9	6.9	11.3	11.3	7.4	7.4	12.0	12.0	
1976	21.2	21.3	21.8	21.9	15.1	15.1	15.6	15.6	
1977	38.1	38.3	48.2	48.4	45.6	45.7	56.0	56.3	
1978	31.1	31.1	34.7	34.8	29.9	30.0	32.6	32.7	
Model D									
(d.f. = 0)		(d.f. = 3)			(d.f. = 0)		(d.f. = 3)		
1975	0.0	0.0	5.0	5.0	0.0	0.0	5.6	5.6	
1976	0.0	0.0	15.3	15.3	0.0	0.0	11.6	11.6	
1977	0.0	0.0	11.5	11.5	0.0	0.0	18.0	18.0	
1978	0.0	0.0	10.2	10.2	0.0	0.0	9.9	9.8	
Model E									
(d.f. = 3)		(d.f. = 6)			(d.f. = 3)		(d.f. = 6)		
1975	7.0	7.0	11.3	11.3	7.3	7.3	12.0	12.0	
1976	21.0	21.1	21.8	21.9	14.8	14.9	15.6	15.6	
1977	33.0	33.0	48.2	48.4	39.5	39.5	56.0	56.3	
1978	32.0	32.1	34.6	34.8	30.9	31.0	32.6	32.7	

Note: $\chi^2_{.99}(3) = 11.34$, $\chi^2_{.99}(4) = 13.28$, $\chi^2_{.99}(5) = 15.09$, $\chi^2_{.99}(6) = 16.81$, $\chi^2_{.99}(7) = 18.48$, and $\chi^2_{.99}(8) = 20.09$.

The fits of Models E-U and E-S are quite similar to those of Models C-U and C-S respectively. This is not surprising since the interpretations of the model are quite similar. Under Model C nonresponse depends on the survey classification when the HH responds while under Model E it depends on the survey classification when the HH does not respond. Since the fits of these two models are similar, we cannot choose between the two models using the data alone. Logically, Model E seems more realistic since we might expect nonresponse to depend on the current victimization status. Since the two models provide similar fits to the data, it may be that the victimization status at the time when the HH responds is generally a good indicator for the victimization status when the HH does not respond. If that is the case, we would prefer to use Model C since it is easier to fit than Model E.

Model A-S, under which nonresponse depends on both the time and on the victimization status when the HH responds fits the 1975, 1976, and 1978 data very well and gives a reasonable fit to the 1977 data. The fits of Model D-S are similar to those of Model A-S with the exception of the 1976 data which is fit much better by Model A-S. Again we cannot choose between Model A and D based on the data alone. (Models A-U and D-U fit the data exactly.) In general, we are quite pleased with the fits of Model A-S to both the number-of-crimes and type-of-crime data from all four years. Since Model A provides a reasonable fit to all the data, we conclude that nonresponse in the NCS does depend on victimization status.

Notice that, in most cases, the fits of the models as measured by X^2 and G^2 do not change much when the symmetric p_{ij} model is used rather than the unconstrained p_{ij} model. Since we gain 3 degrees of freedom going to the more parsimonious, symmetric model for the p_{ij} , we prefer this model to the unconstrained model for the p_{ij} . This choice of the symmetric model for the flow probabilities indicates that there is a certain amount of stability in victimizations reported in the first and second halves of the year in the NCS. This stability comes from the fact that symmetry in the underlying flow probabilities implies equality of marginal totals. Thus, the numbers of HH's having no crimes, one crime, or two or more crimes remain about the same from the first interview of a year to the second year. Similarly, the numbers of HH's having no crimes, a property crime, or a contact crime remain about the same from the first interview of a year to the second year.

5. CONCLUSIONS AND FUTURE WORK

We have seen that the model of symmetry in the matrices of flows among victimization classifications paired with a model under which nonresponse depends on both time and victimization status, provides a good fit to data summaries from the NCS. The same model fits the data when classification of HH's is by number of crimes reported or by type of crime reported.

The work described here is, of course, only an initial attempt to explore nonresponse and flows among victimization classifications in NCS data. For example, we noticed that the estimated symmetric probabilities of flows among the classifications did not appear to change much over the four-year period from 1975 to 1978 but the estimated probabilities of nonresponse did appear to change over this period. One might wish to fit a model to the NCS data which has constant flow probabilities but allows the nonresponse probabilities to change over time. If the nonresponse probabilities do actually change over time, not just from year to year but also from interview period to interview period, then it would be important to try to discover why these probabilities are changing.

In the work presented here, all missing data were treated the same. In fact, data may be missing because a HU rotated out of the sample, because a HH moved into or out of the sampled HU, because no one was at home, because the HH refused to respond, or for some other reason. It may be reasonable to assume that data missing because a HU rotated out of the sample is missing at random, but that other types of nonresponse are not missing at random. Stasny (1988) presents models that allow for different types of nonresponse which could be used with the models of symmetry in flows presented here. In addition, the models here do not allow for HH's which are missing at both interview periods. Since there are, of course, such HH's, one may wish to explore Markov-chain model such as those given in Stasny (1987) which do handle nonresponse at both times.

Most importantly, one may want to consider more natural summaries of the data than were used here. The data used here were summarized by first and second interview for the year. A more meaningful summary would be, say, by month or quarter of the year. If such summaries were used, then the complex nature of the interview schedule for the NCS would have to be considered and accounted for in the models. For example, the response status for a HH would be the same for the six-month reporting period covered at any one interview time. The development of models taking this into account is an important area for future work.

ACKNOWLEDGEMENTS

The National Crime Survey data utilized in this paper were made available by the Inter-University Consortium for Political and Social Research. The data were originally collected by the United States' Law Enforcement Assistance Administration. The longitudinal data set used here was created by the Bureau of Justice Statistics using the quarterly public-use data files. This research was supported in part by a grant from the Bureau of Justice Statistics, U.S. Department of Justice, and the Committee on Law and Justice Statistics, American Statistical Association, which permitted the author to attend two Workshops on the Design and Use of the National Crime Survey. The author takes sole responsibility for the work presented in this paper. The author wishes to thank the referees for their comments on an earlier version of this paper.

APPENDIX I
The Observed Data

			Classification by Number of Victimizations			
			Second Interview			
			Crime Free	Single Crime	Multiple Crime	Missing
1975		Crime Free	1963	256	67	901
First Interview		Single Crime	306	73	31	179
		Multiple Crime	95	26	24	83
		Missing	866	193	91	
1976		Crime Free	1884	257	53	951
First Interview		Single Crime	266	84	24	186
		Multiple Crime	82	34	18	75
		Missing	831	197	106	
1977		Crime Free	1742	260	66	994
First Interview		Single Crime	228	56	31	177
		Multiple Crime	63	31	10	76
		Missing	716	194	79	
1978		Crime Free	1370	157	45	831
First Interview		Single Crime	222	50	14	165
		Multiple Crime	50	18	19	66
		Missing	651	174	57	
			Classification by Type of Crime			
			Second Interview			
			Crime Free	Property Crime	Contact Crime	Missing
1975		Crime Free	1963	271	52	901
First Interview		Property Crime	331	107	22	217
		Contact Crime	70	17	8	45
		Missing	866	225	59	
1976		Crime Free	1884	266	44	951
First Interview		Property Crime	295	111	19	211
		Contact Crime	53	26	4	50
		Missing	831	235	68	
1977		Crime Free	1742	283	43	994
First Interview		Property Crime	262	89	18	194
		Contact Crime	29	12	9	59
		Missing	716	231	42	
1978		Crime Free	1370	173	29	831
First Interview		Property Crime	238	64	14	184
		Contact Crime	34	15	8	47
		Missing	651	184	47	

APPENDIX II: Procedures for Obtaining MLE's of the p and λ Parameters

Note that $x_{..} = \sum_{i=1}^K \sum_{j=1}^K x_{ij}$ is the total number of units responding at both times and $n = x_{..} + x_{.M} + x_{M.}$ is the total sample size. The starting values given below for the iterative procedures are merely suggested values. Other positive values summing to one may be used as initial values for the p -parameter estimates, and other values between zero and one may be used as initial values for the λ -parameter estimates.

MLE's for Unconstrained p_{ij} 's Under Models R, A, B, and C

1. $p_{ij}^{(0)} = x_{ij}/x_{..}$
2. $p_{ij}^{(\nu+1)} = [x_{ij} + x_{iM}p_{ij}^{(\nu)}/p_{i.}^{(\nu)} + x_{MJ}p_{ij}^{(\nu)}/p_{.j}^{(\nu)}]/n$.

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the p_{ij} parameter estimates converge to a desired degree of accuracy.

MLE's for λ 's Under Model A

1. $\lambda_{1j}^{(0)} = x_{M.}/n$ and $\lambda_{2i}^{(0)} = x_{.M}/n$.
2. a) $\lambda_{1j}^{(\nu+1)} = x_{Mj} / \sum_{i=1}^K [x_{ij}/(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)})]$
- b) $\lambda_{2i}^{(\nu+1)} = x_{iM} / \sum_{j=1}^K [x_{ij}/(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)})]$.

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the λ -parameter estimates converge to the desired degree of accuracy. If $x_{hM} > \sum_{j=1}^K x_{hj}$ or $x_{Mh} > \sum_{i=1}^K x_{ih}$ for some h , so that of all units responding in a particular survey classification at one interview time more did not respond at the other interview time than did respond, then the corresponding parameter estimates will, at some step, fall outside of the 0 to 1 range and alternate formulas must be used in place of those given above (see Chen and Fienberg 1974). If for some j $x_{Mj} > \sum_{i=1}^K x_{ij}$, then for that j , step 2a) given above is replaced by

$$\lambda_{1j}^{(\nu+1)} = 1 - \lambda_{2h}^{(\nu)} - (\lambda_{1j}^{(\nu)}/x_{Mj}) \left\{ \sum_{i=1}^K [x_{ij}/(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)})] \right\} (1 - \lambda_{1j}^{(\nu)} - \lambda_{2h}^{(\nu)}),$$

where h is chosen at each step of the iteration so that $\lambda_{2h}^{(\nu)} \geq \lambda_{2i}^{(\nu)}$ for all $i = 1, 2, \dots, K$. If for some i $x_{iM} > \sum_{j=1}^K x_{ij}$, then for that i , step 2b) given above is replaced by

$$\lambda_{2i}^{(\nu+1)} = 1 - \lambda_{1h}^{(\nu)} - (\lambda_{2i}^{(\nu)}/x_{iM}) \left\{ \sum_{j=1}^K [x_{ij}/(1 - \lambda_{1j}^{(\nu)} - \lambda_{2i}^{(\nu)})] \right\} (1 - \lambda_{1h}^{(\nu)} - \lambda_{2i}^{(\nu)}),$$

where h is chosen at each step of the iteration so that $\lambda_{1h}^{(\nu)} \geq \lambda_{1j}^{(\nu)}$ for all $j = 1, 2, \dots, K$.

MLE's for λ 's Under Model B

$$\hat{\lambda}_1 = x_{M\cdot}/n \quad \text{and} \quad \hat{\lambda}_2 = x_{\cdot M}/n.$$

MLE's for λ 's Under Model C

$$1. \lambda_i^{(0)} = (x_{iM} + x_{Mi})/2n.$$

$$2. \lambda_i^{(\nu+1)} = (x_{iM} + x_{Mi}) \left/ \left\{ \sum_{j=1}^K [(x_{ij} + x_{ji}) / (1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)})] \right\} \right.$$

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the λ -parameter estimates converge to the desired degree of accuracy. If $x_{Mi} + x_{iM} > \sum_{j=1}^K (x_{ij} + x_{ji})$ for some i , then as for Model A an alternate formula must be used in place of step 2 above. In such cases, step 2 is replaced by

$$\lambda_i^{(\nu+1)} = 1 - \lambda_h^{(\nu)} - [\lambda_i^{(\nu)} / (x_{iM} + x_{Mi})]$$

$$\left\{ \sum_{j=1}^K [(x_{ij} + x_{ji}) / (1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)})] \right\} (1 - \lambda_h^{(\nu)} - \lambda_i^{(\nu)}),$$

where h is chosen at each step of the iteration so that $\lambda_h^{(\nu)} \geq \lambda_j^{(\nu)}$ for all $j = 1, 2, \dots, K$.

MLE's for Parameters Under Model D-U

$$1. p_{ij}^{(0)} = x_{ij}/x_{\cdot\cdot}, \quad \lambda_{1i}^{(0)} = x_{M\cdot}/n, \quad \text{and} \quad \lambda_{2j}^{(0)} = x_{\cdot M}/n.$$

$$2. p_{ij}^{(\nu+1)} = n^{-1} \left\{ x_{ij} + x_{iM} \left[p_{ij}^{(\nu)} \lambda_{2j}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{2h}^{(\nu)} \right] + x_{Mj} \left[p_{ij}^{(\nu)} \lambda_{1i}^{(\nu)} / \sum_{h=1}^K p_{hj}^{(\nu)} \lambda_{1h}^{(\nu)} \right] \right\}$$

$$\lambda_{1i}^{(\nu+1)} = \sum_{j=1}^K \left[x_{Mj} p_{ij}^{(\nu)} \lambda_{1i}^{(\nu)} / \sum_{h=1}^K p_{hj}^{(\nu)} \lambda_{1h}^{(\nu)} \right] / \sum_{j=1}^K [x_{ij} / (1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)})]$$

$$\lambda_{2j}^{(\nu+1)} = \sum_{i=1}^K \left[x_{iM} p_{ij}^{(\nu)} \lambda_{2j}^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_{2h}^{(\nu)} \right] / \sum_{i=1}^K [x_{ij} / (1 - \lambda_{1i}^{(\nu)} - \lambda_{2j}^{(\nu)})].$$

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the λ -parameter estimates converge to the desired degree of accuracy.

MLE's for Parameters Under Model E-U

$$1. p_{ij}^{(0)} = x_{ij}/x_{..} \quad \text{and} \quad \lambda_i^{(0)} = (x_{M.} + x_{.M})/2n.$$

$$2. p_{ij}^{(\nu+1)} = n^{-1} \left\{ x_{ij} + x_{iM} \left[p_{ij}^{(\nu)} \lambda_j^{(\nu)} / \sum_{h=1}^K p_{ih}^{(\nu)} \lambda_h^{(\nu)} \right] + x_{Mj} \left[p_{ij}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{hj}^{(\nu)} \lambda_h^{(\nu)} \right] \right\}$$

$$\lambda_i^{(\nu+1)} = \left\{ \sum_{j=1}^K x_{jM} \left[p_{ji}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{jh}^{(\nu)} \lambda_h^{(\nu)} \right] + x_{Mj} \left[p_{ij}^{(\nu)} \lambda_i^{(\nu)} / \sum_{h=1}^K p_{hj}^{(\nu)} \lambda_h^{(\nu)} \right] \right\} \\ \times \left\{ \sum_{j=1}^K (x_{ij} + x_{ji}) / (1 - \lambda_i^{(\nu)} - \lambda_j^{(\nu)}) \right\}^{-1}.$$

Step 2 is repeated for $\nu = 0, 1, 2, \dots$ until the λ -parameter estimates converge to the desired degree of accuracy.

REFERENCES

- CHEN, T., and FIENBERG, S.E. (1974). Two-Dimensional Contingency Tables With Both Completely and Partially Cross-Classified Data. *Biometrics*, 30, 629-642.
- DODGE, R.W., and SKOGAN, W.G. (1987). The History, Organization and Utilization of the National Crime Survey. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.
- FIENBERG, S.E. (1980). The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. *The Statistician*, 29, 313-350.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley & Sons.
- MONTAGLIANI, H. (1987). NCS Design. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.
- SAPHIRE, D.G. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey*. Lecture Notes in Statistics, Vol. 23. New York: Springer-Verlag.
- STASNY, E.A. (1986). Estimating Gross Flows Using Panel Data With Nonresponse: An Example From the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- STASNY, E.A. (1987). Some Markov-Chain Models for Nonresponse in Estimating Gross Labor Force Flows. *Journal of Official Statistics*, 3, 359-373.
- STASNY, E.A. (1988). Modeling Non-Ignorable Non-Response in Categorical Panel Data With an Example in Estimating Gross Labor-Force Flows. *Journal of Business and Economic Statistics*, 6, 207-219.
- TAYLOR, B.T. (1987). Redesign of the National Crime Survey. Paper presented at the Workshop on the National Crime Survey, July 6-17, 1987, University of Maryland.
- U.S. DEPARTMENT OF JUSTICE and BUREAU OF JUSTICE STATISTICS (1981). *The National Crime Survey: Working Papers, Volume I: Current and Historical Perspectives*. Washington, DC.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1990:

A. Agresti, <i>University of Florida</i>	N. Laird, <i>Harvard University</i>
J. Armstrong, <i>Statistics Canada</i>	H. Lee, <i>Statistics Canada</i>
W. Bell, <i>U.S. Bureau of the Census</i>	E. Martin, <i>U.S. Bureau of the Census</i>
D. Bellhouse, <i>University of Western Ontario</i>	S.M. Miller, <i>U.S. Bureau of Labor Statistics</i>
K. Bennett, <i>Statistics Canada</i>	A.K. Nigam, <i>Lucknow University</i>
J.M. Berthelot, <i>Statistics Canada</i>	I. Munck, <i>Statistics Sweden</i>
D.A. Binder, <i>Statistics Canada</i>	S. Presser, <i>University of Maryland</i>
K. Bollen, <i>University of North Carolina at Chapel Hill</i>	J.N.K. Rao, <i>Carleton University</i>
P.D. Bourke, <i>University College (Cork)</i>	D.B. Rubin, <i>Harvard University</i>
R. Boyer, <i>Statistics Canada</i>	K. Rust, <i>Westat</i>
R. Carter, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research</i>
G.H. Choudhry, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
N. Cressie, <i>Iowa State University</i>	A. Satin, <i>Statistics Canada</i>
J.-C. Deville, <i>INSEE</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
D. Drew, <i>Statistics Canada</i>	G. Shababb, <i>NPD/Nielsen, Inc.</i>
D.E. Duffy, <i>Bell Communications Research</i>	M. Schenker, <i>UCLA</i>
J.L. Eltinge, <i>Texas A & M University</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
G. Émond, <i>Stone-Consolidated</i>	A. Singh, <i>Statistics Canada</i>
F.J. Fowler, Jr., <i>University of Massachusetts</i>	R. Sitter, <i>Carleton University</i>
W.A. Fuller, <i>Iowa State University</i>	C. Skinner, <i>Iowa State University</i>
J.F. Gosselin, <i>Statistics Canada</i>	K.P. Srinath, <i>Statistics Canada</i>
G. Gray, <i>Statistics Canada</i>	S. Sudman, <i>University of Illinois</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	R. Tiller, <i>U.S. Bureau of Labor Statistics</i>
J. Hox, <i>University of Amsterdam</i>	R.L. Winkler, <i>Duke University</i>
P. Hoyt, <i>Statistics Canada</i>	K. Wolter, <i>A.C. Nielsen</i>
G. Kalton, <i>University of Michigan</i>	C.F.J. Wu, <i>University of Waterloo</i>

Acknowledgements are also due to those who assisted during the production of the 1990 issues: S. Beauchamp and G. Meilleur (Photocomposition), G. Gaulin (Author Services), and M. Haight (Translation Services). Finally we wish to acknowledge J. Clarke, M. Kent, C. Lacroix, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

Survey Methodology

A Journal of Statistical Development and Applications

SUBSCRIBE NOW! ►

Attention Members of Statistical Associations!

- the American Statistical Association
- the Statistical Society of Canada
- the International Association of Survey Statisticians

Members of these associations are eligible for a discount on Survey Methodology. You may subscribe to Survey Methodology when you renew your association membership.



YES! Enter my subscription to Survey Methodology (12-001), two issues per year for only \$35 + 2.45 (7% GST) for a total of \$37.45 in Canada, US\$42 in the U.S., and US\$49 in other countries.

(Please print)

Company _____

Department _____

Attention _____

Address _____

City _____

Province _____

Postal Code _____

Tel. _____

METHOD OF PAYMENT

☐ Purchase Order Number _____

(please enclose)

☐ Payment enclosed \$ _____

☐ Bill me later (max. \$500)

Signature _____

Charge to my:

☐ MasterCard

☐ VISA

Account Number _____

Expiry Date _____

Client Reference Number _____

Cheques or money orders should be made payable to the Receiver General for Canada/Publications. Foreign subscribers please pay in US\$ drawn on a US bank.

PF 03692



Statistics
Canada

Statistique
Canada

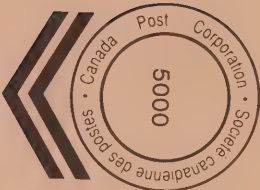
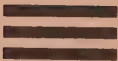
Canada

Business
Reply Mail
**No postage stamp
necessary if mailed
in Canada**

Postage will be paid by
Labour and Household
Surveys Analysis
Division

Courrier-réponse
d'affaires
**Se poste
sans timbre
au Canada**

Le port sera payé par
Division de l'analyse
des enquêtes sur le travail
et les ménages



Statistics Statistique
Canada Canada
Ottawa, Canada
K1A 9Z9



GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

Techniques d'enquête

Une revue sur les méthodes statistiques et leur utilisation

**ABONNEZ-VOUS
DÈS AUJOURD'HUI ! ►**

Êtes-vous membre de l'American Statistical Association, de la Société statistique du Canada ou de l'Association internationale des statisticiens d'enquêtes ?

Si oui, vous avez droit à un rabais sur le prix de l'abonnement. Vous pouvez vous abonner à Techniques d'enquête au moment de payer vos cotisations à votre association.



OUI ! Abonnez-moi à Techniques d'enquête (12-001), deux numéros par année pour seulement 35 \$ + TPS de 7 % (2,45 \$), soit au total 37,45\$ au Canada, 42 \$ US aux États-Unis et 49 \$ US dans les autres pays.

(Écrire en caractères d'imprimerie)

Société _____

Service _____

À l'attention de _____

Adresse _____

Ville _____

Province _____

Code postal _____

Téléphone _____

MODALITÉS DE PAIEMENT

☐ Numéro d'ordre d'achat

(inclure s. v. p.) _____

☐ Paiement inclus \$ _____

☐ Envoyez-moi la facture plus tard

(max. 500 \$)

Signature _____

Numéro de référence du client _____

Le chèque ou mandat-poste doit être fait à l'ordre du Receveur général du Canada — Publications. Les clients à l'étranger paient en \$ US, tirés sur une banque américaine.

PF 03692



Statistique Canada
Statistics Canada

Canada

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1990:

A. Agresti, *University of Florida*
J. Armstrong, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
D. Bellhouse, *University of Western Ontario*
K. Bennett, *Statistique Canada*
J.M. Berthelot, *Statistique Canada*
D.A. Binder, *Statistique Canada*
K. Bollen, *University of North Carolina at Chapel Hill*
P.D. Bourke, *University College (Cork)*
R. Boyer, *Statistique Canada*
R. Carter, *Statistique Canada*
G.H. Choudhry, *Statistique Canada*
N. Cressie, *Iowa State University*
J.-C. Deville, *INSEE*
D. Drew, *Statistique Canada*
D.E. Duffy, *Bell Communications Research*
J.L. Eltinge, *Texas A & M University*
G. Emond, *Stone-Consolidated*
F.J. Fowler, Jr., *University of Massachusetts*
W.A. Fuller, *Iowa State University*
J.F. Gosselin, *Statistique Canada*
G. Gray, *Statistique Canada*
R.M. Groves, *U.S. Bureau of the Census*
J. Hox, *University of Amsterdam*
P. Hoyt, *Statistique Canada*
G. Kalton, *University of Michigan*
C.F.J. Wu, *University of Waterloo*
N. Laird, *Harvard University*
H. Lee, *Statistique Canada*
E. Martin, *U.S. Bureau of the Census*
S.M. Miller, *U.S. Bureau of Labor Statistique*
A.K. Nigam, *Lucknow University*
I. Munck, *Statistics Sweden*
S. Presser, *University of Maryland*
J.N.K. Rao, *Carleton University*
D.B. Rubin, *Harvard University*
K. Rust, *Westat*
I. Sande, *Bell Communications Research*
C.E. Särndal, *Université de Montréal*
A. Satin, *Statistique Canada*
W.L. Schaible, *U.S. Bureau of Labor Statistique*
G. Shabab, *NPD/Nielsen, Inc.*
M. Schenker, *UCLA*
F.J. Scheuren, *U.S. Internal Revenue Service*
A. Singh, *Statistique Canada*
R. Sitter, *Carleton University*
C. Skinner, *Iowa State University*
K.P. Srinath, *Statistique Canada*
S. Sudman, *University of Illinois*
R. Tiller, *U.S. Bureau of Labor Statistique*
R.L. Winkler, *Duke University*
K. Wolter, *A.C. Nielsen*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1990: S. Beauchamp et G. Meilleur (Photocomposition), G. Gaulin (Services aux auteurs) et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à J. Clarke, M. Kent, C. Lacroix, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1990:

REMERCIEMENTS

- A. Agresti, *University of Florida*
J. Armstrong, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
D. Bellhouse, *University of Western Ontario*
K. Bennett, *Statistique Canada*
J. M. Berthelot, *Statistique Canada*
D. A. Binder, *Statistique Canada*
K. Bollen, *University of North Carolina at Chapel Hill*
P. D. Bourke, *University College (Cork)*
R. Boyer, *Statistique Canada*
R. Carter, *Statistique Canada*
I. Sande, *Bell Communications Research*
G. H. Choudhry, *Statistique Canada*
N. Cressie, *Iowa State University*
J.-C. Deville, *INSEE*
D. Drew, *Statistique Canada*
D. E. Duffy, *Bell Communications Research*
J. L. Eltinge, *Texas A & M University*
G. Emond, *Stone-Consolidated*
F. J. Fowler, Jr., *University of Massachusetts*
W. A. Fuller, *Iowa State University*
J. F. Gosselin, *Statistique Canada*
K. P. Srinath, *Statistique Canada*
S. Sudman, *University of Illinois*
R. M. Groves, *U.S. Bureau of the Census*
J. Hox, *University of Amsterdam*
P. Hoyt, *Statistique Canada*
G. Kalton, *University of Michigan*
C. F. J. Wu, *University of Waterloo*
N. Laird, *Harvard University*
H. Lee, *Statistique Canada*
E. Martin, *U.S. Bureau of the Census*
S. M. Miller, *U.S. Bureau of Labor Statistique*
A. K. Nigam, *Lucknow University*
I. Munck, *Statistics Sweden*
S. Presser, *University of Maryland*
J. N. K. Rao, *Carleton University*
D. B. Rubin, *Harvard University*
K. Rust, *Westat*
I. Sande, *Bell Communications Research*
C. E. Särndal, *Université de Montréal*
A. Satin, *Statistique Canada*
W. L. Schaible, *U.S. Bureau of Labor Statistique*
G. Shabbab, *NPD/Nielsen, Inc.*
M. Schenker, *UCLA*
F. J. Scheuren, *U.S. Internal Revenue Service*
A. Singh, *Statistique Canada*
R. Sitter, *Carleton University*
C. Skinner, *Iowa State University*
K. P. Srinath, *Statistique Canada*
S. Sudman, *University of Illinois*
R. Tiller, *U.S. Bureau of Labor Statistique*
R. L. Winkler, *Duke University*
K. Wolter, *A. C. Nielsen*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1990: S. Beauchamp et G. Meilieur (Photocomposition), G. Gaulin (Services aux auteurs) et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à J. Clarke, M. Kent, C. Lacroix, C. Larabie et D. Lermire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

EMV pour les paramètres suivant le modèle E-U

$$1. \quad p_{ij}^{(0)} = x_{ij}/x_{..} \quad \text{et} \quad \lambda_{i(0)}' = (x_{M.} + x_{.M})/2n.$$

$$2. \quad p_{ij}^{(v+1)} = n^{-1} \left\{ x_{ij} + x_{iM} \left[p_{ij}^{(v)} \lambda_{i(v)}^f / \sum_K^{h=1} p_{ih}^{(v)} \lambda_{i(v)}^h \right] + x_{Mj} \left[p_{ij}^{(v)} \lambda_{i(v)}^f / \sum_K^{h=1} p_{hj}^{(v)} \lambda_{i(v)}^h \right] \right\}$$

$$\lambda_{i(v+1)}' = \left\{ \sum_K^{f=1} x_{fM} \left[p_{if}^{(v)} \lambda_{i(v)}^f / \sum_K^{h=1} p_{fh}^{(v)} \lambda_{i(v)}^h \right] + x_{Mj} \left[p_{ij}^{(v)} \lambda_{i(v)}^f / \sum_K^{h=1} p_{hj}^{(v)} \lambda_{i(v)}^h \right] \right\}$$

$$\times \left\{ \sum_K^{f=1} (x_{ij} + x_{jf}) / (1 - \lambda_{i(v)}^f - \lambda_{j(v)}^f) \right\}^{-1}.$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$ jusqu'à ce que les estimations des paramètres λ convergent vers le degré d'exactitude prescrit.

BIBLIOGRAPHIE

CHEN, T., et FIENBERG, S.E. (1974). Two-Dimensional Contingency Tables With Both Completely and Partially Cross-Classified Data. *Biometrics*, 30, 629-642.
 DODGE, R.W., et SKOGAN, W.G. (1987). The History, Organization and Utilization of the National Crime Survey. Mémoire présenté au Workshop on the National Crime Survey, du 6 au 17 juillet 1987, University of Maryland.

FIENBERG, S.E. (1980). The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey. *The Statistician*, 29, 313-350.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley & Sons.

MONTAGLIANI, H. (1987). NCS Design. Mémoire présenté au Workshop on the National Crime Survey, du 6 au 17 juillet 1987, University of Maryland.

SAPHIRE, D.G. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey*. Lecture Notes in Statistics, Vol. 23. New York: Springer-Verlag.

STASNY, E.A. (1986). Estimating Gross Flows Using Panel Data With Nonresponse: An Example From the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A. (1987). Some Markov-Chain Models for Nonresponse in Estimating Gross Labor Force Flows. *Journal of Official Statistics*, 3, 359-373.

STASNY, E.A. (1988). Modeling Non-Ignorable Non-Response in Categorical Panel Data With an Example in Estimating Gross Labor-Force Flows. *Journal of Business and Economic Statistics*, 6, 207-219.

TAYLOR, B.T. (1987). Redesign of the National Crime Survey. Mémoire présenté au Workshop on the National Crime Survey, du 6 au 17 juillet 1987, University of Maryland.

U.S. DEPARTMENT OF JUSTICE et BUREAU OF JUSTICE STATISTICS (1981). *The National Crime Survey: Working Papers, Volume I: Current and Historical Perspectives*. Washington, DC.

EMV pour les λ suivant le modèle B

$$\lambda_1 = x_{M\cdot}/n \quad \text{et} \quad \lambda_2 = x_{M\cdot}/n.$$

EMV pour les λ suivant le modèle C

$$1. \lambda_{(0)}^i = (x_{iM} + x_{Mi})/2n.$$

$$2. \lambda_{(v+1)}^i = (x_{iM} + x_{Mi}) \bigg/ \left\{ \sum_{k=1}^f [(x_{iU} + x_{iI})/(1 - \lambda_{(v)}^i - \lambda_{(v)}^f)] \right\}.$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$ jusqu'à ce que les estimations des paramètres λ convergent vers le degré d'exactitude prescrit. Si $x_{Mi} + x_{iM} > \sum_{j=1}^K (x_{iU} + x_{ij})$ pour certaines valeurs de i , il faut, tout comme dans le cas du modèle A, utiliser une autre formule à la place de l'étape 2 ci-devant. En pareil cas, on remplace l'étape 2 par

$$\lambda_{(v+1)}^i = 1 - \lambda_{(v)}^h - [\lambda_{(v)}^i/(x_{iM} + x_{Mi})]$$

$$\left\{ \sum_{k=1}^f [(x_{iU} + x_{iI})/(1 - \lambda_{(v)}^i - \lambda_{(v)}^f)] \right\} (1 - \lambda_{(v)}^h - \lambda_{(v)}^i),$$

où h est choisi à chaque étape de la méthode itérative de façon à ce que $\lambda_{(v)}^h \geq \lambda_{(v)}^f$ pour tous les $f = 1, 2, \dots, K$.

EMV pour les paramètres suivant le modèle D-U

$$1. d_{(0)}^{ij} = x_{ij}/x_{\cdot\cdot}, \quad \lambda_{(0)}^{i1} = x_{M\cdot}/n, \quad \text{et} \quad \lambda_{(0)}^{2f} = x_{M\cdot}/n.$$

$$2. d_{(v+1)}^{ij} = u^{-1} \left\{ x_{ij} + x_{iM} \left[d_{(v)}^{ij} \lambda_{(v)}^{2f} \bigg/ \sum_{k=1}^h d_{(v)}^{ik} \lambda_{(v)}^{2h} \right] + x_{Mj} \left[d_{(v)}^{ij} \lambda_{(v)}^{i1} \bigg/ \sum_{k=1}^h d_{(v)}^{ik} \lambda_{(v)}^{h1} \right] \right\}$$

$$\lambda_{(v+1)}^{i1} = \sum_{k=1}^f \left[x_{Mj} d_{(v)}^{ij} \lambda_{(v)}^{i1} \bigg/ \sum_{k=1}^h d_{(v)}^{ik} \lambda_{(v)}^{h1} \right] (1 - \lambda_{(v)}^{i1} - \lambda_{(v)}^{2f})$$

$$\lambda_{(v+1)}^{2f} = \sum_{k=1}^f \left[x_{iM} d_{(v)}^{ij} \lambda_{(v)}^{2f} \bigg/ \sum_{k=1}^h d_{(v)}^{ik} \lambda_{(v)}^{2h} \right] (1 - \lambda_{(v)}^{i1} - \lambda_{(v)}^{2f}).$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$ jusqu'à ce que les estimations des paramètres λ convergent vers le degré d'exactitude prescrit.

ANNEXE II: Méthodes d'obtention des EMV des paramètres p et λ

Notons que $x_{..} = \sum_K^i \sum_{f=1}^K x_{ij}$ est le nombre total d'unités participant aux deux cycles d'interview et $n = x_{..} + x_M$ est la taille de l'échantillon global. Les valeurs de départ utilisées ci-dessous le sont uniquement à titre indicatif. On peut utiliser comme valeurs initiales des estimations des paramètres p toutes autres valeurs positives dont la somme est égale à 1 et, comme valeurs initiales pour les paramètres λ , toutes autres valeurs situées entre zéro et un.

EMV pour les p_{ij} non assujetties à des contraintes suivant les modèles R, A, B et C

$$1. \quad p_{ij(0)} = x_{ij}/x_{..}$$

$$2. \quad p_{(v+1)}^{ij} = [x_{ij} + x_M p_{(v)}^{ij} / p_{(v)}^{i.} + x_M p_{(v)}^{.j} / p_{(v)}^{..}] / n.$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$ jusqu'à ce que les estimations des paramètres p_{ij} convergent vers le degré d'exactitude prescrit.

EMV pour les λ suivant le modèle A

$$1. \quad \lambda_{(0)}^{1f} = x_M/n \quad \text{et} \quad \lambda_{(0)}^{2f} = x_{..M}/n.$$

$$2. \quad \text{a) } \lambda_{(v+1)}^{1f} = x_M / \left[\sum_K^i [x_{ij} / (1 - \lambda_{(v)}^{1f} - \lambda_{(v)}^{2f})] \right]$$

$$\text{b) } \lambda_{(v+1)}^{2f} = x_M / \left[\sum_K^i [x_{ij} / (1 - \lambda_{(v)}^{1f} - \lambda_{(v)}^{2f})] \right].$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$ jusqu'à ce que les estimations des paramètres λ convergent vers le degré d'exactitude prescrit. Si $x_{hM} > \sum_{f=1}^K x_{hf}$ ou $x_{Mh} > \sum_{i=1}^K x_{ih}$ pour certaines valeurs de h , de telle sorte que, parmi toutes les unités ayant été classées dans une catégorie donnée à l'occasion d'un cycle d'interview, le nombre d'unités n'ayant pas participé à l'autre cycle d'interview est plus élevé que celui des unités y ayant participé, alors la valeur des estimations des paramètres correspondants se situera à l'extérieur de la plage des valeurs prescrites (0 à 1) à l'une ou l'autre étape et il faudra utiliser d'autres formules à la place de celles exposées ci-dessus (voir Chen et Fienberg 1974). Si pour certaines valeurs de j $x_{Mj} > \sum_{i=1}^K x_{ij}$, il faut, pour ces valeurs, remplacer l'étape 2a) ci-dessus par

$$\lambda_{(v+1)}^{1f} = 1 - \lambda_{(v)}^{2h} - \left(\lambda_{(v)}^{1f} / x_{Mj} \right) \left\{ \sum_K^i [x_{ij} / (1 - \lambda_{(v)}^{1f} - \lambda_{(v)}^{2f})] \right\} (1 - \lambda_{(v)}^{1f} - \lambda_{(v)}^{2h}),$$

où h est choisi à chaque étape de la méthode itérative de façon à ce que $\lambda_{(v)}^{2h} \geq \lambda_{(v)}^{2f}$ pour tous les $i = 1, 2, \dots, K$. Si pour certaines valeurs de i , $x_{iM} > \sum_{f=1}^K x_{if}$, il faut, pour ces valeurs, remplacer l'étape 2b) ci-dessus par

$$\lambda_{(v+1)}^{2f} = 1 - \lambda_{(v)}^{1h} - \left(\lambda_{(v)}^{2f} / x_{iM} \right) \left\{ \sum_K^i [x_{ij} / (1 - \lambda_{(v)}^{1f} - \lambda_{(v)}^{2f})] \right\} (1 - \lambda_{(v)}^{1h} - \lambda_{(v)}^{2f}),$$

où h est choisi à chaque étape de la méthode de façon à ce que $\lambda_{(v)}^{1h} \geq \lambda_{(v)}^{1f}$ pour tous les $j = 1, 2, \dots, K$.

ANNEXE I

Les observations

Classement selon le nombre d'actes criminels déclarés			
Deuxième interview			
Aucun crime	Un seul crime	Crimes multiples	Non-réponse

Classement selon le genre d'acte criminel déclaré			
Deuxième interview			
1975 - Première interview	Aucun crime 1963	Un crime 73	31
	Crimes multiples 306	26	179
	Non-réponse 95	83	901
1976 - Première interview	Aucun crime 1884	Un crime 257	53
	Crimes multiples 266	84	24
	Non-réponse 82	34	186
1977 - Première interview	Aucun crime 1742	Un crime 260	66
	Crimes multiples 228	56	31
	Non-réponse 63	31	10
1978 - Première interview	Aucun crime 1370	Un crime 157	45
	Crimes multiples 222	50	14
	Non-réponse 50	18	19
		174	57

1975 - Première interview	Aucun crime 1963	Crime contre la propriété 331	52
	Crime contre la personne 70	107	22
	Non-réponse 866	17	8
1976 - Première interview	Aucun crime 1884	Crime contre la propriété 266	44
	Crime contre la personne 295	111	19
	Non-réponse 53	26	4
1977 - Première interview	Aucun crime 1742	Crime contre la propriété 283	43
	Crime contre la personne 262	89	18
	Non-réponse 716	12	9
1978 - Première interview	Aucun crime 1370	Crime contre la propriété 238	29
	Crime contre la personne 34	64	14
	Non-réponse 651	15	8
		184	47

Dans le présent article, toutes les données manquantes ont été traitées de la même façon. De fait, les données peuvent être manquantes parce qu'une UL a été supprimée de l'échantillon par renouvellement, parce qu'un ménage a emménagé dans l'UL sélectionnée ou l'a quittée, parce que personne n'était sur place, parce que le ménage a refusé de répondre ou pour quel-que autre raison. On peut raisonnablement supposer que la non-réponse découle du fait qu'une UL a été supprimée de l'échantillon par renouvellement constitue une non-réponse aléatoire, mais que les autres types de non-réponse ne sont pas aléatoires. À cet égard, Stasny (1988) présente des modèles pour différents types de non-réponse qui peuvent être combinés aux modèles décrivant la symétrie des flux présentés dans cet article. En outre, les modèles que nous avons étudiés ne permettent pas de tenir compte des ménages qui ne participent à aucun des deux cycles d'interview. Comme il existe sûrement de tels ménages, il pourrait être intéressant d'étudier les modèles markoviens, comme ceux qu'on trouve dans Stasny (1987), qui permettent de tenir compte de la non-réponse aux deux cycles d'interview.

Surtout, il serait sûrement utile d'étudier des sommaires de données plus naturels que ceux auxquels nous avons eu recours. Alors que nous avons utilisé des sommaires portant sur chaque cycle d'interview, il serait plus révélateur d'utiliser des sommaires mensuels ou trimestriels. En pareil cas, il faudrait prendre en considération la nature complexe du calendrier d'interviews de la NCS et en tenir compte dans les modèles. Ainsi, le code de réponse attribué à un ménage resterait inchangé pour la totalité de la période de référence de six mois sur laquelle porte une interview. L'élaboration de modèles tenant compte de ce paramètre constitue à n'en pas douter une piste de recherche qu'il importe d'explorer.

REMERCIEMENTS

Les données de la National Crime Survey sur lesquelles se fonde le présent article ont été mises à notre disposition par le Consortium interuniversitaire pour la recherche politique et sociale. Ces données ont été initialement recueillies par la United States Law Enforcement Assistance Administration. L'ensemble de données longitudinales utilisé a été créé par le Bureau of Justice Statistics, à partir des fichiers de données trimestrielles à grande diffusion. La présente recherche a été rendue possible en partie grâce à une subvention du Bureau of Justice Statistics, U.S. Department of Justice, et du Committee on Law and Justice Statistics, American Statistical Association, qui a permis à l'auteur d'assister aux deux ateliers sur la conception et l'utilisation de la National Crime Survey. Les opinions exprimées dans le cadre du présent article sont entièrement celles de l'auteur. L'auteur remercie sincèrement les membres du comité de lecture pour leurs commentaires relatifs à une version antérieure du mémoire.

il nous est impossible d'opter en faveur de l'un ou de l'autre en nous fondant uniquement sur les observations. Logiquement, le modèle E semble plus réaliste, puisque nous pouvons nous attendre à ce que la non-réponse varie selon que le ménage a été ou non victime d'actes criminels. Comme les deux modèles s'ajustent aux données de façon semblable, il se peut que l'état (victime ou non-victime) dans lequel se trouve le ménage au moment où il répond constitue, en règle générale, un bon indicateur de l'état dans lequel il se trouve au moment où il ne répond pas. Si cette hypothèse se confirmait, nous préfererions utiliser le modèle C puisqu'il est plus facile à ajuster que le modèle E.

Le modèle A-S, suivant lequel la non-réponse est fonction du moment où l'interview a lieu et de l'état dans lequel se trouve le ménage lorsqu'il répond, s'ajuste très bien aux données de 1975, 1976 et 1978 et raisonnablement bien aux données de 1977. L'ajustement du modèle D-S est semblable à celui du modèle A-S, sauf dans le cas des données de 1976 auxquelles il s'ajuste beaucoup mieux que le modèle A-S. Encore une fois, il nous est impossible d'opter pour le modèle A ou le modèle D en nous fondant uniquement sur les observations. (Les modèles A-U et D-U s'ajustent parfaitement aux données.) De façon générale, nous sommes très satisfaits de l'ajustement du modèle A-S tant aux données sur le nombre d'actes criminels qu'aux données sur le genre d'acte criminel, et ce pour les quatre années considérées. Comme le modèle A s'ajuste raisonnablement bien à toutes les données, nous devons conclure que la non-réponse observée dans le cadre de la NCS varie selon que le ménage a été ou non victime d'actes criminels.

Il convient de noter que, dans la majorité des cas, l'ajustement des modèles mesuré à l'aide de X^2 et G^2 varie très peu lorsqu'on utilise le modèle des p_{ij} symétriques plutôt que celui des p_{ij} non assujetties à des contraintes. Comme le modèle symétrique, plus parcimonieux, nous permet de gagner 3 degrés de liberté, nous le préférons au modèle non assujéti à des contraintes. Le fait que nous choisissons le modèle symétrique pour les probabilités de flux indique que le nombre d'actes criminels déclarés dans le cadre des deux cycles annuels d'interview de la NCS est relativement stable. Cette stabilité découle du fait que la symétrie des probabilités de flux sous-jacentes implique l'égalité des totaux marginaux. Aussi, le nombre de ménages ne déclarant aucun crime, déclarant un crime ou déclarant plus d'un crime demeure à peu près le même du premier au deuxième cycle d'interview d'une année. Il en va de même du nombre de ménages n'ayant été victimes d'aucun crime, ayant été victimes d'un crime contre la propriété ou ayant été victimes d'un crime contre la personne.

5. CONCLUSIONS ET PISTES DE RECHERCHE

Nous avons vu que, une fois combiné avec un modèle suivant lequel la non-réponse est fonction à la fois de la valeur de t et de l'état (victime ou non-victime) dans lequel se trouve le ménage, le modèle décrivant la symétrie des matrices de flux dans les catégories d'actes criminels déclarés s'ajuste bien aux sommaires de données de la NCS. De plus, ce modèle s'ajuste aux données peu importe que les ménages soient classés selon le nombre d'actes criminels déclarés ou selon le genre d'acte criminel déclaré.

Bien sûr, les présents travaux ne constituent qu'une tentative initiale d'étudier la non-réponse et les flux dans les catégories d'actes criminels déclarés dans le cadre de la NCS. Ainsi, nous avons souligné que les probabilités estimées de flux symétriques dans les catégories n'ont pas semblé fluctuer beaucoup au cours de la période de quatre ans allant de 1975 à 1978, tandis que les probabilités estimées de non-réponse ont semblé afficher une certaine variation. Il pourrait se révéler intéressant d'ajuster aux données de la NCS un modèle décrivant à la fois des probabilités de flux constantes et des probabilités de non-réponse variant en fonction du moment où l'interview a eu lieu, non seulement d'une année à l'autre mais d'un cycle d'interview à l'autre, il est essentiel d'essayer de découvrir la cause de ces fluctuations.

Tableau 11

Ajustement des modèles

Classement selon le nombre d'actes criminels déclarés		Classement selon le genre d'acte criminel déclaré	
p_{ij} non assujetties à des contraintes		p_{ij} non assujetties à des contraintes	
X^2	G^2	X^2	G^2
Modèle R			
(d.l. = 5)	(d.l. = 8)	(d.l. = 5)	(d.l. = 8)
1975 42.7	45.9	1975 42.0	42.0
1976 70.2	69.7	1976 56.9	53.8
1977 74.2	83.9	1977 68.4	65.4
1978 61.7	64.9	1978 47.6	50.1
Modèle A			
(d.l. = 0)	(d.l. = 3)	(d.l. = 0)	(d.l. = 3)
1975 0.0	4.4	1975 0.0	4.6
1976 0.0	0.6	1976 0.0	0.5
1977 0.0	10.1	1977 0.0	10.5
1978 0.0	3.7	1978 0.0	2.7
Modèle B			
(d.l. = 4)	(d.l. = 7)	(d.l. = 4)	(d.l. = 7)
1975 42.7	45.9	1975 42.0	41.5
1976 69.1	68.5	1976 56.9	53.8
1977 47.1	58.7	1977 68.4	65.4
1978 47.6	50.1	1978 47.4	50.1
Modèle C			
(d.l. = 3)	(d.l. = 6)	(d.l. = 3)	(d.l. = 6)
1975 6.9	11.3	1975 7.4	12.0
1976 21.2	21.8	1976 15.1	15.6
1977 38.1	48.2	1977 56.0	56.3
1978 31.1	34.7	1978 30.0	32.6
Modèle D			
(d.l. = 0)	(d.l. = 3)	(d.l. = 0)	(d.l. = 3)
1975 0.0	5.0	1975 0.0	5.6
1976 0.0	15.3	1976 0.0	11.6
1977 0.0	11.5	1977 0.0	18.0
1978 0.0	10.2	1978 0.0	9.8
Modèle E			
(d.l. = 3)	(d.l. = 6)	(d.l. = 3)	(d.l. = 6)
1975 7.0	11.3	1975 7.3	12.0
1976 21.0	21.8	1976 14.9	15.6
1977 33.0	48.2	1977 39.5	56.3
1978 32.0	34.6	1978 31.0	32.7

Nota: $\chi^2_{.99}(3) = 11.34$, $\chi^2_{.99}(4) = 13.28$, $\chi^2_{.99}(5) = 15.09$, $\chi^2_{.99}(6) = 16.81$, $\chi^2_{.99}(7) = 18.48$, et $\chi^2_{.99}(8) = 20.09$.

cycles d'interview. Lorsque le modèle C est combiné avec le modèle symétrique pour les paramètres p , nous obtenons des fréquences prévues de cellule symétriques pour les flux d'observations. Notons que, selon les observations montrées dans l'annexe 1, le nombre de cas de non-réponse est beaucoup plus élevé au deuxième qu'au premier cycle d'interview des années 1977 et 1978. C'est en raison de cet écart entre les taux de non-réponse que le modèle C s'ajuste moins bien aux données de 1977 et de 1978.

L'ajustement des modèles E-U et E-S à l'échantillon d'observations est presque identique à celui des modèles C-U et C-S respectivement. Cela n'a rien de surprenant quand on sait que les interprétations des modèles sont assez semblables. Suivant le modèle C, la non-réponse est fonction de la catégorie de réponse dans laquelle est classé le ménage lorsqu'il répond, tandis que suivant le modèle E, elle est fonction de la catégorie de réponse dans laquelle est classé le ménage lorsqu'il ne répond pas. Comme l'ajustement de ces deux modèles est semblable,

Tableau 10

Estimations de λ_j suivant le modèle E

Classement selon le nombre d'actes criminels déclarés		Classement selon le genre d'acte criminel déclaré	
λ_1	λ_2	λ_1	λ_3

p_{ij} non assujettis à des contraintes

1975	.202	.285	.348	.201	.302	.336
1976	.211	.275	.387	.209	.286	.419
1977	.210	.315	.372	.209	.318	.394
1978	.224	.340	.342	.225	.326	.385
	(.0065)	(.0208)	(.0296)	(.0065)	(.0203)	(.0333)

p_{ij} symétriques

1975	.202	.285	.351	.201	.301	.341
1976	.211	.274	.389	.209	.287	.418
1977	.213	.301	.376	.213	.309	.391
1978	.224	.343	.338	.225	.329	.379
	(.0065)	(.0204)	(.0298)	(.0065)	(.0199)	(.0339)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

4.4 Ajustement des modèles

Le tableau 11 indique les valeurs de X^2 et G^2 ainsi que le nombre de degrés de liberté s'y rapportant pour l'ensemble des douze modèles (y compris le modèle D-U qui doit s'ajuster partialement aux données) et pour les deux types de classement des données. On notera que les modèles ont été ajustés pour illustrer les méthodes élaborées dans le présent article et que nous n'avons pas tenu compte de toute la complexité du plan de sondage. Bien que l'existence de grappes ne pose aucun problème dans notre sous-échantillon de données de la NCS, nous préférons, dans une analyse plus approfondie, ajuster les modèles aux données des diverses strates séparément puis combiner les estimations relatives au strates afin d'obtenir des estimations pour l'ensemble de la population.

À l'évidence, ni le modèle R, modèle pour la non-réponse aléatoire, ni le modèle B, selon lequel la probabilité de non-réponse est fonction uniquement de la valeur de t , ne s'ajuste bien aux données que ce soit selon le modèle des p_{ij} non assujettis à des contraintes ou selon le modèle symétrique des p_{ij} .

Les modèles C-U et C-S s'ajustent assez bien aux données de 1975 et raisonnablement bien aux données de 1976. Comme le modèle C-S s'ajuste raisonnablement bien aux données et qu'il est plus parcimonieux, nous le préférons au modèle C-U. Selon le modèle C, la probabilité de non-réponse est fonction uniquement de la catégorie dans laquelle l'unité est classée au moment de l'interview à laquelle le ménage a répondu et non du moment où l'interview a eu lieu. Aussi le modèle C est-il le modèle de symétrie des probabilités de non-réponse dans le cadre des deux

4.3 Estimations des paramètres suivant les modèles D et E

Comme ils exigent que tous les paramètres soient estimés simultanément, les modèles D et E sont plus difficiles à ajuster que les modèles R, A, B et C. Quel que soit l'ensemble de données de la NCS considéré, les fonctions de vraisemblance suivant les modèles D et E n'étaient pas classiques et il a fallu effectuer une recherche par quadrillage portant sur les valeurs possibles des paramètres λ afin de déterminer des points de départ appropriés pour la méthode itérative. Comme la recherche par quadrillage portant sur les six paramètres λ suivant le modèle D nécessitait beaucoup de temps, nous avons obtenu les estimations des paramètres selon le modèle D-S mais non selon le modèle D-U. Les estimations des paramètres p selon le modèle D-S sont présentées au tableau 3b pour le classement selon le nombre d'actes criminels déclarés et au tableau 4b pour le classement selon le genre d'acte criminel déclaré. Les estimations des paramètres λ suivant le modèle D-S sont présentées au tableau 9 pour les deux types de classement. Les estimations des paramètres p suivant les modèles E-U et E-S sont présentées au tableau 3c pour le classement selon le nombre d'actes criminels déclarés et au tableau 4c pour le classement selon le genre d'acte criminel déclaré. Enfin, les estimations des paramètres λ suivant les modèles E-U et E-S sont présentées au tableau 10 pour les deux types de classement. Il convient de noter que, suivant les modèles D et E, la valeur estimée de p_{11} , la probabilité qu'une unité reste classée dans la catégorie "aucun crime", est légèrement inférieure à la valeur correspondante suivant les modèles R, A, B et C; les estimations des autres paramètres p suivant les modèles D et E sont toutefois légèrement supérieures aux estimations correspondantes suivant les modèles R, A, B et C. Selon les modèles D et E, la valeur estimative du paramètre λ , la probabilité estimée de non-réponse, s'accroît généralement en fonction du nombre d'actes criminels déclarés ou de la gravité de l'acte criminel déclaré. Lorsque la valeur des estimations décroît en fonction du nombre d'actes criminels déclarés ou de la gravité de l'acte criminel déclaré (pour les données de 1978 sur le nombre d'actes criminels déclarés suivant le modèle E-S), la diminution est faible et inférieure à l'erreur type estimée des estimations.

Tableau 9

Estimations de λ_{1i} et de λ_{2j} suivant le modèle D-S

Classement selon le nombre d'actes criminels déclarés						Classement selon le genre d'acte criminel déclaré			
λ_{11}	λ_{12}	λ_{13}	λ_{21}	λ_{22}	λ_{23}	λ_{11}	λ_{12}	λ_{13}	λ_{23}

1975	.210	.246	.319	.194	.321	.387	.208	.264	.319	.192	.339	.372
	(.0085)	(.0303)	(.0368)	(.0085)	(.0282)	(.0362)	(.0084)	(.0249)	(.0523)	(.0085)	(.0235)	(.0507)
1976	.204	.276	.339	.217	.273	.444	.203	.280	.383	.215	.297	.453
	(.0083)	(.0274)	(.0344)	(.0084)	(.0291)	(.0331)	(.0083)	(.0244)	(.0443)	(.0084)	(.0255)	(.0416)
1977	.175	.307	.380	.249	.298	.374	.175	.304	.438	.248	.315	.341
	(.0086)	(.0301)	(.0403)	(.0089)	(.0326)	(.0439)	(.0086)	(.0243)	(.0424)	(.0089)	(.0259)	(.0491)
1978	.211	.278	.290	.236	.413	.384	.211	.276	.293	.236	.411	.391
	(.0094)	(.0282)	(.0433)	(.0099)	(.0261)	(.0443)	(.0094)	(.0264)	(.0563)	(.0098)	(.0246)	(.0567)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Estimations des λ_{1j} et des λ_{2i} suivant le modèle A

Classement selon le nombre d'actes criminels déclarés		Classement selon le genre d'acte criminel déclaré	
λ_{11}	λ_{12}	λ_{13}	λ_{23}
1975	.208 (.0062)	.272 (.0159)	.246 (.0303)
1976	.206* (.0063)	.261* (.0152)	.253 (.0319)
	.267* (.0268)	.397* (.0066)	.285 (.0319)
	.254* (.0153)	.236* (.0066)	
	.267* (.0248)		
	.206 (.0063)		
	.278 (.0146)		
	.381 (.0327)		
	.235 (.0066)		
	.220 (.0064)		
	.322 (.0321)		
	.280 (.0151)		
	.208 (.0062)		
	.275 (.0242)		
	.234 (.0147)		
	.281 (.0171)		
	.326 (.0285)		
	.321* (.0300)		
	.280* (.0176)		
	.207* (.0072)		
	.305* (.0174)		
	.275 (.0144)		
	.267 (.0327)		
	.258 (.0069)		
	.269* (.0079)		
	.343* (.0364)		
	.280* (.0166)		
	.334* (.0362)		
1978	.207* (.0072)	.316* (.0182)	.417 (.0369)
1977	.192 (.0064)	.263 (.0152)	.269 (.0159)

Notes: L'astérisque * indique les cas où la fonction de vraisemblance n'est pas classique. Les erreurs types estimées sont indiquées entre parenthèses.

Estimations de λ_1 et λ_2 suivant le modèle B

Classement selon le nombre d'actes criminels déclarés ou selon le genre d'acte criminel déclaré	
λ_1	λ_2
.223 (.0058)	.226 (.0058)
.225 (.0059)	.240 (.0060)
.209 (.0059)	.264 (.0064)
.227 (.0067)	.273 (.0071)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Estimations des λ_i suivant le modèle C

Classement selon le nombre d'actes criminels déclarés		Classement selon le genre d'acte criminel déclaré	
λ_1	λ_2	λ_1	λ_2
1975	.214 (.0039)	.214 (.0039)	.262 (.0109)
1976	.221 (.0040)	.221 (.0040)	.266 (.0109)
1977	.225 (.0041)	.225 (.0041)	.273* (.0115)
1978	.237* (.0046)	.237* (.0046)	.292* (.0130)
	.252 (.0118)	.214 (.0039)	.284 (.0262)
	.300 (.0199)	.300 (.0199)	.333 (.0289)
	.271 (.0126)	.271 (.0126)	.339* (.0286)
	.312* (.0236)	.312* (.0236)	.339* (.0299)

Notes: L'astérisque * indique les cas où la fonction de vraisemblance n'est pas classique. Les erreurs types estimées sont indiquées entre parenthèses.

4.2 Estimations des paramètres λ selon les modèles R, A, B et C

Le lecteur se souviendra que les estimations des paramètres λ selon les modèles R, A, B et C sont identiques, peu importe qu'on utilise le modèle non assujéti à des contraintes ou le modèle symétrique pour les paramètres p . Suivant la méthode itérative servant à estimer les paramètres λ selon les modèles A et C, le critère de convergence utilisé était que les estimations des paramètres λ varient d'au plus 0.0005 d'une étape à l'autre. Lorsque la convergence s'est produite après moins de 10,000 étapes, elle a nécessité de 41 à 4150 étapes à partir des estimations initiales des paramètres suggérées à l'annexe II. Dans certains cas, les facteurs de la fonction de vraisemblance pour les observations portant uniquement sur les paramètres λ n'étaient pas classiques. Cela est particulièrement vrai dans le cas des fonctions de vraisemblance pour les données de 1978 suivant les modèles A et C. En pareil cas, on a eu recours à une recherche par quadrillage pour déterminer les points de départ appropriés du processus itératif. Nous avons également effectué une recherche grossière par quadrillage dans tous les cas afin de vérifier si, lorsque la méthode itérative convergeait, elle semblait avoir convergé vers un maximum global plutôt que vers un maximum local.

Les estimations des paramètres λ établies pour le classement selon le nombre d'actes criminels déclarés et le classement selon le genre d'acte criminel déclaré suivant les modèles R, A, B et C sont présentées respectivement dans les tableaux 5, 6, 7 et 8.

Il convient de noter que, suivant les modèles R et B, les estimations des paramètres λ sont les mêmes pour les deux types de classement (selon le nombre d'actes criminels déclarés et selon le genre d'acte criminel déclaré), puisque la probabilité de non-réponse selon ces deux modèles n'est pas fonction de la catégorie de réponse. Suivant les modèles A et C, les estimations des paramètres λ correspondant à la catégorie "aucun crime" sont les mêmes, compte tenu de l'erreur d'arrondi, pour les deux types de classement (selon le nombre d'actes criminels déclarés et selon le genre d'acte criminel déclaré), puisque les ménages n'ayant pas été victimes de crime sont les mêmes dans les deux cas. On notera également que, suivant les modèles A et C, la valeur estimative du paramètre λ , la probabilité estimée d'être un non-répondant, s'accroît généralement en fonction du nombre d'actes criminels déclarés ou de la gravité de l'acte criminel déclaré.

Tableau 5
Estimations des λ suivant le modèle R

Données classées selon le nombre d'actes criminels déclarés ou selon le genre d'acte criminel déclaré		λ
1975	.224 (.0035)	
1976	.232 (.0035)	
1977	.237 (.0036)	
1978	.250 (.0040)	

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Tableau 4c
Estimations des p_{ij} pour les flux dans les catégories de genre
d'actes criminels déclaré suivant les modèles E-U et E-S

Modèle non assujéti à des contraintes	Deuxième interview				Modèle symétrique
	Aucun crime	Crime contre la propriété	Crime contre la personne	Aucun crime	Crime contre la propriété

1975	Aucun crime	Crime contre la propriété	Crime contre la personne	Aucun crime	Crime contre la propriété
	.636 (.0100)	.124 (.0063)	.027 (.0033)	.636 (.0101)	.052 (.0047)
	.111 (.0062)	.053 (.0047)	.009 (.0020)	.026 (.0026)	.011 (.0016)
	.024 (.0034)	.012 (.0023)	.005 (.0016)	.117 (.0046)	.005 (.0016)

1976	Aucun crime	Crime contre la propriété	Crime contre la personne	Aucun crime	Crime contre la propriété
	.641 (.0098)	.110 (.0059)	.024 (.0033)	.641 (.0098)	.052 (.0048)
	.110 (.0065)	.051 (.0048)	.016 (.0028)	.110 (.0046)	.015 (.0021)
	.028 (.0041)	.014 (.0028)	.005 (.0019)	.110 (.0046)	.005 (.0019)

1977	Aucun crime	Crime contre la propriété	Crime contre la personne	Aucun crime	Crime contre la propriété
	.636 (.0108)	.107 (.0060)	.015 (.0028)	.641 (.0105)	.049 (.0048)
	.138 (.0076)	.050 (.0048)	.011 (.0027)	.121 (.0051)	.011 (.0018)
	.023 (.0035)	.010 (.0022)	.009 (.0023)	.121 (.0105)	.009 (.0022)

1978	Aucun crime	Crime contre la propriété	Crime contre la personne	Aucun crime	Crime contre la propriété
	.641 (.0117)	.124 (.0071)	.020 (.0033)	.640 (.0117)	.048 (.0054)
	.111 (.0078)	.048 (.0055)	.014 (.0031)	.118 (.0056)	.013 (.0021)
	.022 (.0040)	.012 (.0029)	.009 (.0025)	.118 (.0117)	.008 (.0025)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Tableau 4b
Estimations des p_{ij} pour les flux dans les catégories de genre
d'acte criminel suivant le modèle D-S

Modèle symétrique		
Deuxième interview		
Aucun crime	Un seul crime	Crimes multiples

1975	Première	Aucun crime	.635	.118	.026
			(.0101)	(.0046)	(.0026)
			.118	.052	.011
	interview	Crime contre la propriété	.118	(.0046)	(.0016)
		Crime contre la personne	.026	.011	.005
			(.0026)	(.0016)	(.0016)

1976	Première	Aucun crime	.641	.110	.026
			(.0098)	(.0046)	(.0028)
			.110	.052	.015
	interview	Crime contre la propriété	.110	(.0048)	(.0021)
		Crime contre la personne	.026	.015	.004
			(.0028)	(.0021)	(.0019)

1977	Première	Aucun crime	.642	.120	.019
			(.0104)	(.0052)	(.0024)
			.120	.050	.011
	interview	Crime contre la propriété	.120	(.0049)	(.0019)
		Crime contre la personne	.019	.011	.008
			(.0024)	(.0019)	(.0022)

1978	Première	Aucun crime	.636	.121	.020
			(.0117)	(.0057)	(.0025)
			.121	.049	.012
	interview	Crime contre la propriété	.121	(.0055)	(.0021)
		Crime contre la personne	.020	.012	.008
			(.0025)	(.0021)	(.0025)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Tableau 4a
Estimations des p_j pour les flux dans les catégories de genre
d'actes criminels suivant les modèles, R, A, B, et C

Modèle	Modèle non assujéti à des contraintes		Deuxième interview	
	Crime contre la personne	Crime contre la propriété	Crime contre la personne	Crime contre la propriété

1975	Aucun crime	.666 (.0075)	.105 (.0053)	.022 (.0026)	.666 (.0075)	.111 (.0037)	.024 (.0018)
	Crime contre la propriété	.118 (.0054)	.044 (.0038)	.010 (.0019)	.111 (.0037)	.044 (.0038)	.008 (.0013)
	Crime contre la personne	.025 (.0026)	.007 (.0016)	.004 (.0012)	.024 (.0018)	.008 (.0013)	.004 (.0012)

1976	Aucun crime	.669 (.0076)	.108 (.0055)	.023 (.0028)	.669 (.0021)	.108 (.0011)	.022 (.0010)
	Crime contre la propriété	.108 (.0053)	.047 (.0040)	.010 (.0021)	.108 (.0011)	.047 (.0019)	.011 (.0009)
	Crime contre la personne	.021 (.0025)	.012 (.0021)	.002 (.0011)	.022 (.0010)	.011 (.0009)	.002 (.0012)

1977	Aucun crime	.670 (.0079)	.128 (.0061)	.019 (.0026)	.671 (.0078)	.115 (.0039)	.018 (.0018)
	Crime contre la propriété	.103 (.0053)	.041 (.0039)	.008 (.0018)	.115 (.0039)	.041 (.0040)	.008 (.0014)
	Crime contre la personne	.016 (.0025)	.008 (.0021)	.006 (.0018)	.018 (.0018)	.008 (.0014)	.006 (.0017)

1978	Aucun crime	.671 (.0087)	.104 (.0064)	.019 (.0031)	.671 (.0088)	.112 (.0044)	.019 (.0021)
	Crime contre la propriété	.119 (.0063)	.040 (.0044)	.010 (.0024)	.112 (.0044)	.040 (.0044)	.010 (.0017)
	Crime contre la personne	.019 (.0029)	.011 (.0025)	.006 (.0020)	.019 (.0021)	.010 (.0017)	.006 (.0020)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Il convient de remarquer, dans les tableaux 3a et 4a, que les matrices de flux des probabilités estimées selon le modèle non assujéti à des contraintes pour les paramètres p_{ij} semblent passablement symétriques. Le modèle pour la symétrie des flux se révèle donc un modèle dont on peut raisonnablement envisager l'utilisation. Notons également que les estimations des paramètres p_{ij} ne semblent pas varier beaucoup au cours de la période de quatre ans. Les ajustements de ces deux modèles pour les paramètres p_{ij} seront étudiés pour chacun des quatre modèles de non-réponse au paragraphe 4.4 ci-après.

Tableau 3c

Estimations des p_{ij} pour les flux dans les catégories de nombre d'actes criminels suivant les modèles E-U et E-S

Modèle non assujéti à des contraintes	Deuxième interview		
	Aucun crime	Un seul crime	Crimes multiples
Modèle symétrique	Aucun crime	Un seul crime	Crimes multiples

1975	Première	Aucun crime	.639	.102	.031	.639	.106	.035
		Un seul crime	.110	.033	.016	.106	.033	.015
		Crimes multiples	.039	.014	.016	.035	.015	.016
			(.0039)	(.0025)	(.0027)	(.0028)	(.0019)	(.0027)
	Première	Aucun crime	.645	.103	.032	.645	.101	.033
		Un seul crime	.098	.037	.017	.101	.037	.017
		Crimes multiples	.035	.017	.016	.033	.017	.016
			(.0037)	(.0027)	(.0029)	(.0029)	(.0021)	(.0029)
1976	Première	Aucun crime	.636	.124	.037	.642	.106	.033
		Un seul crime	.094	.031	.021	.106	.030	.020
		Crimes multiples	.029	.020	.008	.033	.020	.008
			(.0036)	(.0031)	(.0024)	(.0033)	(.0023)	(.0025)
	Première	Aucun crime	.636	.124	.037	.642	.106	.033
		Un seul crime	.094	.031	.021	.106	.030	.020
		Crimes multiples	.029	.020	.008	.033	.020	.008
			(.0036)	(.0031)	(.0024)	(.0033)	(.0023)	(.0025)
1977	Première	Aucun crime	.636	.124	.037	.642	.106	.033
		Un seul crime	.094	.031	.021	.106	.030	.020
		Crimes multiples	.029	.020	.008	.033	.020	.008
			(.0036)	(.0031)	(.0024)	(.0033)	(.0023)	(.0025)
	Première	Aucun crime	.639	.106	.029	.637	.112	.028
		Un seul crime	.117	.041	.011	.112	.041	.013
		Crimes multiples	.027	.016	.015	.028	.013	.015
			(.0037)	(.0032)	(.0030)	(.0029)	(.0021)	(.0030)
1978	Première	Aucun crime	.639	.106	.029	.637	.112	.028
		Un seul crime	.117	.041	.011	.112	.041	.013
		Crimes multiples	.027	.016	.015	.028	.013	.015
			(.0037)	(.0032)	(.0030)	(.0029)	(.0021)	(.0030)
	Première	Aucun crime	.639	.106	.029	.637	.112	.028
		Un seul crime	.117	.041	.011	.112	.041	.013
		Crimes multiples	.027	.016	.015	.028	.013	.015
			(.0037)	(.0032)	(.0030)	(.0029)	(.0021)	(.0030)
	Première	Aucun crime	.639	.106	.029	.637	.112	.028
		Un seul crime	.117	.041	.011	.112	.041	.013
		Crimes multiples	.027	.016	.015	.028	.013	.015
			(.0037)	(.0032)	(.0030)	(.0029)	(.0021)	(.0030)
	Première	Aucun crime	.639	.106	.029	.637	.112	.028
		Un seul crime	.117	.041	.011	.112	.041	.013
		Crimes multiples	.027	.016	.015	.028	.013	.015
			(.0037)	(.0032)	(.0030)	(.0029)	(.0021)	(.0030)

Nota: Les erreurs types estimées son indiquées entre parenthèses.

Dans tous les cas, il y a rapidement eu convergence, après au plus six étapes. Les estimations des paramètres p_{ij} lorsque les ménages sont classés selon le nombre d'actes criminels déclarés sont présentées au tableau 3a, tant pour le modèle non assujéti à des contraintes que pour le modèle symétrique. On trouve au tableau 4a les estimations correspondantes lorsque les ménages sont classés selon le genre d'acte criminel déclaré.

Tableau 3b

Estimations des p_{ij} pour les flux dans les catégories de nombre d'actes criminels suivant le modèle D-S

Modèle symétrique		
Deuxième interview		
Crimes multiples	Un seul crime	Aucun crime

1975	Aucun crime	.638	(.0104)	.106	(.0047)	.035	(.0029)
	Un seul crime	.106	(.0047)	.033	(.0039)	.015	(.0019)
	Crimes multiples	.035	(.0029)	.015	(.0019)	.016	(.0027)

1976	Aucun crime	.645	(.0100)	.100	(.0045)	.034	(.0029)
	Un seul crime	.100	(.0045)	.037	(.0041)	.017	(.0021)
	Crimes multiples	.034	(.0029)	.017	(.0021)	.015	(.0029)

1977	Aucun crime	.642	(.0109)	.106	(.0054)	.033	(.0032)
	Un seul crime	.106	(.0054)	.031	(.0043)	.021	(.0023)
	Crimes multiples	.033	(.0032)	.021	(.0023)	.009	(.0025)

1978	Aucun crime	.636	(.0118)	.114	(.0056)	.028	(.0029)
	Un seul crime	.114	(.0056)	.040	(.0051)	.013	(.0021)
	Crimes multiples	.028	(.0029)	.013	(.0021)	.015	(.0030)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

Tableau 3a
Estimations des p_{ij} pour les flux dans les catégories de nombre d'actes criminels suivant les modèles R, A, B, et C

Modèle	Modèle non assujéti à des contraintes		Deuxième interview	
	Aucun crime	Un seul crime	Aucun crime	Un seul crime
Crimes multiples	Crimes multiples	Crimes multiples	Crimes multiples	Crimes multiples

1975	Aucun crime	.666 (.0075)	.098 (.0050)	.029 (.0031)	.666 (.0075)	.102 (.0035)	.032 (.0022)
	Un seul crime	.106 (.0051)	.029 (.0031)	.014 (.0023)	.102 (.0035)	.029 (.0031)	.012 (.0015)
	Crimes multiples	.036 (.0032)	.011 (.0021)	.012 (.0021)	.032 (.0022)	.012 (.0015)	.012 (.0021)

1976	Aucun crime	.669 (.0076)	.101 (.0052)	.029 (.0033)	.669 (.0076)	.099 (.0036)	.030 (.0022)
	Un seul crime	.098 (.0051)	.034 (.0034)	.014 (.0025)	.099 (.0036)	.034 (.0034)	.014 (.0017)
	Crimes multiples	.031 (.0030)	.014 (.0023)	.011 (.0022)	.030 (.0022)	.014 (.0017)	.010 (.0022)

1977	Aucun crime	.670 (.0079)	.115 (.0058)	.032 (.0034)	.671 (.0079)	.103 (.0037)	.030 (.0023)
	Un seul crime	.092 (.0051)	.026 (.0032)	.016 (.0026)	.103 (.0037)	.026 (.0032)	.016 (.0018)
	Crimes multiples	.028 (.0030)	.016 (.0026)	.006 (.0017)	.030 (.0023)	.016 (.0018)	.006 (.0017)

1978	Aucun crime	.671 (.0087)	.097 (.0062)	.027 (.0035)	.671 (.0087)	.105 (.0043)	.027 (.0025)
	Un seul crime	.111 (.0061)	.032 (.0040)	.009 (.0022)	.105 (.0043)	.032 (.0040)	.010 (.0017)
	Crimes multiples	.027 (.0034)	.013 (.0027)	.013 (.0026)	.027 (.0025)	.010 (.0017)	.013 (.0026)

Nota: Les erreurs types estimées sont indiquées entre parenthèses.

4.1 Estimations des paramètres p selon les modèles R, A, B et C

Selon les modèles R, A, B et C, les estimations des paramètres p comme on l'a déjà dit, ne dépendent pas du mécanisme de non-réponse. Suivant la méthode itérative servant à estimer le paramètre p_{ij} selon le modèle non assujéti à des contraintes et selon le modèle symétrique, le critère utilisé pour mettre fin à l'itération était que les chiffres prévus dans la cellule (i, j) de la matrice de flux, np_{ij} , varient d'au plus 0.5 d'une étape à l'autre du processus d'itération.

Méthode itérative d'estimation des paramètres suivant le modèle E-S

$$1. \; p_{(0)}^{ii} = x_{ii}/x_{..}$$

$$p_{(0)}^{ij} = (x_{ij} + x_{ji})/2x_{..} \quad \text{pour } i \neq j$$

$$\lambda_{(0)}^i = (x_{M.} + x_{.M})/2n.$$

$$2. \; p_{(v+1)}^{ii} = n^{-1} \left\{ x_{ii} + (x_{iM} + x_{Mi}) \left[p_{(v)}^{ii} \lambda_{(v)}^i / \sum_K^h d_{(v)}^{ih} \lambda_{(v)}^h \right] \right\}$$

$$p_{(v+1)}^{ij} = (2n)^{-1} \left\{ x_{ij} + x_{ji} + (x_{iM} + x_{Mi}) \left[d_{(v)}^{ij} \lambda_{(v)}^j / \sum_K^h d_{(v)}^{ih} \lambda_{(v)}^h \right] \right\}$$

$$+ (x_{iM} + x_{Mj}) \left[d_{(v)}^{ij} \lambda_{(v)}^i / \sum_K^h d_{(v)}^{jh} \lambda_{(v)}^h \right] \quad \text{pour } i \neq j$$

$$\lambda_{(v+1)}^i = \sum_K^f \left[(x_{iM} + x_{Mj}) d_{(v)}^{ji} \lambda_{(v)}^j / \sum_K^h d_{(v)}^{jh} \lambda_{(v)}^h \right]$$

$$/ \sum_K^f \left[(x_{ij} + x_{ji}) / (1 - \lambda_{(v)}^i - \lambda_{(v)}^j) \right].$$

On répète l'étape 2 pour $v = 0, 1, 2, \dots$, jusqu'à ce que les estimations des paramètres convergent vers le degré d'exactitude prescrit. Les estimations initiales données à l'étape 1 le sont uniquement à titre indicatif. Il est possible d'utiliser toute autre valeur située entre zéro et un et satisfaisant à la contrainte selon laquelle la somme des p_{ij} doit être égale à un.

4. AJUSTEMENT DES MODELES AUX DONNEES DE LA NCS

Les modèles décrits à la section 3 ont été ajustés aux données de la NCS dont il est question à la section 2. Le lecteur se souviendra que les données de la NCS recueillies pour chaque année de 1975 à 1978 sont classées selon le nombre d'actes criminels déclarés lors de chacune des deux entrevues réalisées au cours de l'année et selon le genre d'acte criminel. Comme nous avons utilisé trois catégories de réponses, nous avons $K = 3$. Les erreurs types des estimations des paramètres ont été calculées à partir de la matrice des observations.

$$\left\{ \left[\frac{{}^q 1 \chi_{(a)}^q d}{\sum_X} \bigg/ \frac{{}^1 1 \chi_{(a)}^1 d} \right] {}^1 W_X + \left[\frac{{}^q 2 \chi_{(a)}^q d}{\sum_X} \bigg/ \frac{{}^1 2 \chi_{(a)}^1 d} \right] {}^1 W_X + {}^1 X \right\}_{1-u} = (1+a) {}^1 d \cdot 2$$

$$\left[\begin{matrix} {}^{\mathcal{U}}\chi_{(a)}^{\mathcal{U}} d \\ \sum_K \left[\begin{matrix} {}^{\mathcal{U}}\chi_{(a)}^{\mathcal{U}} d \\ \end{matrix} \right] \end{matrix} \right] w_X + {}^{\mathcal{U}}\chi_X + {}^{\mathcal{U}}\chi_X \Big\} _{1-} (u\mathcal{Z}) = {}_{(1+a)}^{\mathcal{U}} d$$

$$f \neq 1 \mod \left\{ \left[\binom{y_1}{(a)} \chi_{(a)}^{y_1} d^{\sum_{X=1}^q} / \binom{l_1}{(a)} \chi_{(a)}^{l_1} d^{\sum_X} \right] W_X + \right.$$

$$\left[\binom{f}{(a)} \chi - \binom{f}{(a)} \chi - 1 \right] / \binom{f}{X} \Big/ \left[\binom{q}{(a)} \chi \binom{q}{(a)} d \right] \Big/ \left[\binom{f}{(a)} \chi \binom{f}{(a)} d \right] \Big/ \left[\binom{f}{X} \right] = \binom{f}{(1+a)} \chi$$

$$\cdot \left[\binom{f_Z}{(a)} \chi - \binom{I}{(a)} \chi - 1 \right] / \binom{I}{X} \sum_X \bigg/ \left[\binom{Y_Z}{(a)} \chi \binom{Y}{(a)} d \sum_X \bigg/ \left[\binom{f_Z}{(a)} \chi \binom{f}{(a)} d^{W_X} \right] \sum_X \right] = \binom{f_Z}{(1+a)} \chi$$

On répète l'étape 2 pour $\nu = 0, 1, 2, \dots$, jusqu'à ce que les estimations des paramètres convergent vers le degré d'exactitude prescrit. Les estimations initiales données à l'étape 1 le sont uniquement à titre indicatif. Il est possible d'utiliser toute autre valeur située entre zéro et un et satisfaisant à la contrainte selon laquelle la somme des p_{ij} doit être égale à un.

3.5 Estimation des paramètres p et λ suivant le modèle E

Selon le modèle E-U ou le modèle E-S, les fonctions de vraisemblance pour les observations ne peuvent se décomposer en facteurs et les estimations de tous les paramètres doivent être obtenues simultanément. Stasny (1988) expose une méthode itérative permettant d'obtenir les EMV suivant le modèle E-U. On trouve les équations relatives à cette méthode à l'annexe II. Suivant le modèle E-S, la fonction de vraisemblance pour les observations est la suivante:

$$\left\{ \hat{n}_x [({}^f\chi - {}^l\chi - 1)] \prod_K {}^1{}_K^f \prod_K {}^1{}_K^l \right\} \times \left\{ \hat{n}_x^l d \prod_{-1}^1 {}^1{}_K^f \prod_K {}^1{}_K^l \right\} \times \left\{ \hat{n}_x^l d \prod_K {}^{1+l}{}_K^f \prod_K {}^1{}_K^l \right\} \times \left\{ \hat{n}_x^l d \prod_K {}^1{}_K^l \right\}$$

$$(\mathcal{E}) \quad \cdot \left\{ \chi_{M^x} \left[\chi_{!d}^{\perp} \prod_K^{\perp=1} \right] \prod_K^{\perp=1} \right\} \times \left\{ \chi_{M^x} \left[\chi_{!d}^{\perp} \prod_K^{\perp=1} \right] \prod_K^{\perp=1} \right\} \times$$

On maximise l'équation (3) en tenant compte de la condition que la somme des p_{ij} est égale à un. En général, il faut avoir recours à une méthode itérative pour obtenir les EMV. La méthode itérative utilisée aux fins de l'analyse des données dont il est fait état dans la section 4 est la suivante.

suiivante.

$$\cdot \left\{ \prod_K \chi_{\mathcal{W}_X} \right\}^{\prod_{f=1}^I} \times \left\{ \prod_K \chi_{\mathcal{W}_X} \right\}^{\prod_{f=1}^I} \times \left\{ \prod_K \chi_{\mathcal{W}_X} \right\}^{\prod_{f=1}^I} \prod_K \chi_{\mathcal{W}_X}^{\prod_{f=1}^I}$$
$$\hat{\chi} = (x_M + x_{M'})/2n.$$

Selon le modèle D-U ou le modèle D-S, les fonctions de vraisemblance pour les observations ne peuvent se décomposer en facteurs et les estimations de tous les paramètres doivent être obtenues simultanément. On trouve dans Stasny (1988) une méthode itérative permettant d'obtenir les EMV suivant le modèle D-U. Les équations relatives à cette méthode sont présentées à l'annexe II. Suivant le modèle D-S, la fonction de vraisemblance pour les observations est la suivante:

$$(2) \quad \left\{ \prod_K \left[\prod_{l=1}^f \prod_{i=1}^f \lambda_{li}^{f_l} \right] \right\}_{W_f} \times \left\{ \prod_K \left[\prod_{l=1}^f \prod_{i=1}^f d_{li}^{f_l} \right] \right\}_{M_f} \times$$

Méthode itérative d'estimation des paramètres suivant le modèle D-S

$$\cdot x/!!x = {}_{(0)}!!d \cdot 1$$

$$d_{(0)}^{ij} = (x_{ij} + x_{ji})/2x.. \text{ pour } i \neq j$$

$$u/\cdot W_X = {}_{(0)}^I \mathbb{V}$$

$$u/W \cdot x = {}_{(0)}^f \chi$$

distributions asymptotiques du χ^2 avec le nombre de degrés de liberté indiqué ci-devant, pour autant que le modèle soit correct. Nous utiliserons ci-après la notation "modèle R-U" pour désigner la combinaison du modèle R pour les paramètres λ et du modèle non assujéti aux contraintes pour les p_{ij} , ainsi que la notation "modèle R-S" pour désigner la combinaison du modèle R pour les paramètres λ et du modèle symétrique pour les p_{ij} . Nous ferons appel à une notation similaire pour désigner la combinaison des modèles A, B, C, D ou E pour les paramètres λ avec un des deux modèles pour les p_{ij} .

3.3 Estimation des paramètres p et λ suivant les modèles R, A, B et C

Les fonctions de vraisemblance pour les huit modèles créés en combinant un des deux modèles pour les p_{ij} au modèle R, A, B ou C pour les λ_{ij} se décomposent en deux facteurs: le premier est une fonction des paramètres p uniquement et le second, une fonction des paramètres λ uniquement. On peut donc obtenir les EMV séparément pour les deux ensembles de paramètres. En outre, les estimations des paramètres p ne varient pas en fonction du modèle utilisé pour les paramètres λ et les estimations des paramètres λ sont indépendantes du modèle utilisé pour les paramètres p .

Chen et Fienberg (1974) exposent une méthode itérative permettant d'obtenir les EMV pour les paramètres p suivant le modèle non assujéti à des contraintes combiné au modèle R, A, B ou C pour les paramètres λ . On trouve les équations relatives à cette méthode à l'annexe II. Suivant le modèle symétrique pour les paramètres p combiné au modèle R, A, B ou C pour les paramètres λ , le facteur de la fonction de vraisemblance portant uniquement sur les p_{ij} se lit comme suit:

$$\left\{ \prod_{k=1}^K d_{x_{ij}}^{n_{ij}} \right\} \times \left\{ \prod_{k=1}^K \prod_{j=i+1}^K d_{x_{ij}}^{n_{ij}} \right\} \times \left\{ \prod_{k=1}^K \prod_{j=1}^{i-1} d_{x_{ij}}^{n_{ij}} \right\} \quad (1)$$

où le point substitué à l'indice inférieur indique que la sommation s'étend à toutes les valeurs de cet indice. On maximise l'équation (1) en tenant compte de la condition que la somme des p_{ij} est égale à un. En général, il faut utiliser une méthode itérative pour obtenir les EMV. Soit $x_{..} = \sum_{i=1}^K \sum_{j=1}^K x_{ij}$ le nombre total d'unités observées dans les deux cycles d'interview et $n = x_{..} + x_M + x_M$, le nombre total d'unités observées dans au moins un des deux cycles d'interview. Alors la méthode itérative utilisée aux fins de l'analyse des données dont il est fait état dans la section 4 est la suivante.

Méthode itérative d'estimation des p_{ij} symétriques suivant les modèles R, A, B et C

$$1. \quad p_{ij}^{(0)} = x_{ij} / x_{..}$$

$$p_{ij}^{(0)} = (x_{ij} + x_{ji}) / 2x_{..} \quad \text{pour } i \neq j.$$

$$2. \quad p_{ij}^{(v+1)} = [x_{ij} + (x_{iM} + x_{Mi}) d_{(v)}^{ij} / d_{(v)}^{ii}] / n$$

$$p_{ij}^{(v+1)} = [(x_{ij} + x_{ji}) + (x_{iM} + x_{Mi}) d_{(v)}^{ij} / d_{(v)}^{ii} + (x_{jM} + x_{Mj}) d_{(v)}^{ij} / d_{(v)}^{jj}] / 2n \quad \text{pour } i \neq j.$$

des flux symétriques, $p_{ij} = p_{ji}$ pour $i \neq j$ de telle sorte que la probabilité qu'une unité passe de l'état i au temps 1 à l'état j au temps 2 est la même que la probabilité qu'une unité passe de l'état j au temps 1 à l'état i au temps 2. Notons que la symétrie des probabilités de cellules de la matrice de flux implique une égalité entre les totaux marginaux de ligne et de colonne. Aussi le modèle décrivant la symétrie des probabilités de flux implique-t-il une certaine stabilité de la population puisque le nombre prévu d'unités classées dans une catégorie de réponses donnée est le même au temps 1 et au temps 2.

Selon la définition ci-dessus, λ_{1ij} , la probabilité que les unités classées dans les catégories de réponses i au temps 1 et j au temps 2 soient manquantes au temps t est fonction du moment auquel la non-réponse se produit et des catégories de réponses dans lesquelles l'unité est classée au temps 1 et au temps 2. Nous considérons, pour estimer ces probabilités, six modèles plus simples que nous exposons ci-après avec les degrés de liberté correspondants suivant les deux modèles pour les p_{ij} :

d.l. p_{ij} non assujetties à des contraintes		d.l. p_{ij} symétriques	
Modèle R: $\lambda_{1ij} = \lambda$,	$2K - 1$	$(K^2 + 3K - 2)/2$	
Modèle A: $\lambda_{1ij} = \lambda_{1j}$, $\lambda_{2ij} = \lambda_{2j}$	0	$(K^2 - - K)/2$	
Modèle B: $\lambda_{1ij} = \lambda_i$,	$2K - 2$	$(K^2 + 3K - 4)/2$	
Modèle C: $\lambda_{1ij} = \lambda_j$, $\lambda_{2ij} = \lambda_j$,	K	$(K^2 + K)/2$	
Modèle D: $\lambda_{1i} = \lambda_{1j}$, $\lambda_{2ij} = \lambda_{2j}$,	0	$(K^2 - - K)/2$	
Modèle E: $\lambda_{1ij} = \lambda_i$, $\lambda_{2ij} = \lambda_j$,	K	$(K^2 + K)/2$	

Le modèle R est le modèle pour la non-réponse aléatoire. Suivant ce modèle, il n'y a qu'une seule probabilité de non-réponse pour toutes les unités aux deux temps considérés, quelle que soit la catégorie de réponse. Suivant le modèle A, la probabilité qu'une unité soit manquante au temps t est fonction de la valeur de t et de la catégorie de réponses dans laquelle l'unité est classée au moment où elle répond. Notons que si on utilise le modèle A pour les paramètres λ et le modèle non assujéti à des contraintes pour les p_{ij} , alors le modèle est un modèle saturé qui s'ajuste parfaitement aux données. Suivant le modèle B, la probabilité qu'une unité soit manquante au temps t est fonction uniquement de la catégorie de réponses dans laquelle l'unité est classée au moment où elle répond. Suivant le modèle D, la probabilité qu'une unité soit manquante au temps t est fonction à la fois de la valeur de t et de la catégorie de réponses dans laquelle l'unité est classée au moment où elle est manquante. Si on utilise le modèle D pour les paramètres et le modèle non assujéti à des contraintes pour les p_{ij} , alors le modèle est un modèle saturé qui s'ajuste parfaitement aux données. Suivant le modèle E, la probabilité qu'une unité soit manquante au temps t est fonction uniquement de la catégorie de réponses dans laquelle l'unité est classée au moment où elle est manquante. Suivant le modèle R, la non-réponse est dite parfaitement aléatoire. Suivant les modèles A, B et C, on n'a pas à tenir compte de la non-réponse puisque le mécanisme de non-réponse est uniquement fonction des observations. Suivant les modèles D et E, il faut tenir compte de la non-réponse puisque le mécanisme de non-réponse est fonction des données manquantes. Pour plus de renseignements sur les types de non-réponse, voir Little et Rubin (1987).

Nous décrivons dans les deux prochaines sous-sections les méthodes d'ajustement des modèles présentés ci-dessus. On peut évaluer l'ajustement des modèles à l'aide de la statistique X^2 de Pearson ou du rapport des vraisemblances G^2 . Les deux statistiques suivent des

Nous supposons que chaque unité se trouverait dans une des cellules de la matrice $K \times K$ des catégories de réponses si elle était observée au cours des deux cycles d'interview. Définissons p_{ij} comme la probabilité qu'une unité passe de l'état i au temps 1 à l'état j au temps 2, où i et j prennent les valeurs 1, 2, ..., K . Chaque unité de la cellule (i, j) de la matrice de catégories de réponses a une chance d'être manquante à l'occasion d'un des deux cycles d'interview. Définissons λ_{ij} comme la probabilité qu'une des unités de la cellule (i, j) ne réponde pas au temps t et soit donc classée comme manquante. Alors, les probabilités se rattachant aux observations

Tableau 2
Probabilités se rattachant aux observations

Temps		Temps 2	
1	2	1	2
\vdots	K	$\{ (1 - \lambda_{11} - \lambda_{21}) p_{11} \}$	$\left\{ \sum_{j=1}^K p_{1j} \lambda_{2j} \right\}$
		Manquante	Manquante

Si on suppose que les probabilités p_{ij} obéissent à une loi multinomiale, la fonction de vraisemblance pour les observations est proportionnelle à

$$\left\{ \prod_{i=1}^K \prod_{j=1}^K p_{ij} (1 - \lambda_{1ij} - \lambda_{2ij}) \right\}^{x_{ij}}$$
$$\times \left\{ \prod_{j=1}^K \left[\sum_{i=1}^K p_{ij} \lambda_{2ij} \right]^{x_{2j}} \right\}$$
$$\times \left\{ \prod_{i=1}^K \left[\sum_{j=1}^K p_{ij} \lambda_{1ij} \right]^{x_{1i}} \right\}$$

Le nombre des paramètres libres définis ci-dessus est égal à $3K^2 + 2K - 1$, tandis que le nombre de cellules d'observations avec une seule contrainte s'appliquant à la taille de l'échantillon global est seulement de $K^2 + 2K$. Le nombre de paramètres à estimer à l'aide des observations est donc trop élevé et il nous faut réduire le nombre des paramètres compris dans le modèle. Dans la suite de l'article, nous procédons à une telle réduction en considérant deux modèles pour les paramètres p_{ij} et six modèles pour les paramètres λ_{ij} .

3.2 Modèles pour les probabilités p et λ

Nous considérons deux modèles pour estimer les probabilités (p_{ij}) de flux au sein des catégories de réponses: le modèle non assujéti à des contraintes et le modèle des flux symétriques. Suivant le modèle des probabilités de flux non assujéti à des contraintes, il existe une probabilité différente, p_{ij} , pour chaque cellule (i, j) de la matrice de flux. Suivant le modèle

Le premier jeu illustre la répartition des ménages selon le nombre d'actes criminels déclarés au cours de chacune des deux interviews réalisées chaque année. Les catégories utilisées sont: aucun crime (aucun acte criminel déclaré), un seul crime (un acte criminel déclaré), crimes multiples (au moins deux actes criminels déclarés) et unités manquantes (le ménage n'a pas répondu ou a été supprimé de l'échantillon par renouvellement). Le deuxième jeu de matrices illustre la répartition des ménages selon le genre de crime déclaré. Les catégories utilisées sont: aucun crime, crime contre la propriété (vol avec effraction, vol simple et vol de véhicule à moteur), crime contre la personne (viol, voies de fait, vol qualifié, vol de sac à l'arraché et vol à la tire) et non-réponse. Ces catégories de genre de crime sont les mêmes que celles utilisées dans le cadre de la NCS. Lorsqu'un même ménage a déclaré de multiples crimes, on utilise le crime le plus grave déclaré aux fins du classement (les crimes contre la personne sont jugés plus graves que les crimes contre la propriété).

On notera le nombre élevé de cas de non-réponse dans les matrices illustrées à l'annexe I. Seulement 50% environ des ménages ayant participé à au moins une des deux interviews se sont prêtés aux deux cycles d'interview. Les modèles présents dans la prochaine section nous permettront de tenir compte de cette non-réponse tout en étudiant la structure de la matrice sous-jacente des probabilités de flux dans les catégories d'actes criminels déclarés.

3. LES MODELES

Nous examinerons dans la présente section la forme générale des modèles qui seront utilisés pour étudier les flux bruts dans les catégories d'actes criminels déclarés dans le cadre de la NCS. Les modèles épousent la forme proposée par Chen et Fienberg (1974) pour les tableaux de contingence comportant des données à classification complète ou partielle. Les modèles pour la non-réponse et le modèle pour la non-réponse aléatoire ont été élaborés par Stasny (1986). Toutefois, le modèle pour la symétrie des flux ne figure pas dans les ouvrages antérieurs de l'auteur. Nous avons choisi de présenter les modèles sous forme générale parce qu'ils peuvent être appliqués à des problèmes autres que l'estimation des flux bruts au sein des catégories d'actes criminels déclarés dans le cadre de la NCS.

3.1 Modèle pour les observations

Considérons des unités d'observation se prêtant à au moins un des deux cycles d'interview d'une enquête. Supposons que, lorsqu'une unité participe à l'enquête, elle est classée dans une de K catégories. Lorsqu'une unité ne participe pas à l'enquête, elle est classée comme manquante. Alors on peut représenter le flux des observations d'une interview à l'autre comme dans le tableau 1.

Tableau 1

Sommaire des observations

Temps 2		Temps 1	
1	2	1	2
x_{11}	x_{12}	x_{11}	x_{12}
x_{21}	x_{22}	x_{21}	x_{22}
x_{K1}	x_{K2}	x_{K1}	x_{K2}
...		...	
x_{KM}		x_{KM}	

Manquante

le ménage (vol avec effraction, vol simple et vol de véhicule à moteur). Ces questions sont suivies de onze questions de filtrage visant à obtenir de chaque membre du ménage des renseignements sur les crimes contre la personne dont il a été victime (voies de fait, viol et vol qualifié). Enfin, un rapport d'acte criminel est établi pour chaque crime déclaré en réponse aux questions de filtrage.

Pour obtenir de plus amples renseignements sur le plan de sondage et l'historique de la NCS, on peut consulter en outre les ouvrages du U.S. Department of Justice and Bureau of Justice Statistics (1981), Saphire (1984), Dodge et Skogan (1987) et Montagliani (1987). La NCS utilise un nouveau plan de sondage depuis 1986. On trouvera une description du remaniement de la NCS ainsi que des travaux de recherche s'y rapportant dans Taylor (1987). Toutefois, les données utilisées dans cet article ont été recueillies suivant le plan de sondage initial de la NCS.

2.2 Les données longitudinales

Les données utilisées dans cet article sont tirées d'un ensemble de données longitudinales de grande taille comprenant tous les renseignements recueillis par l'intermédiaire des interviews régulières de la NCS réalisées de janvier 1975 à juin 1979, à l'exception des interviews réalisées auprès des UL introduites dans l'échantillon par renouvellement en 1979. Afin que les données soient plus simples à traiter, les présents travaux portent uniquement sur un sous-ensemble de données. Ce sous-ensemble a été établi en prenant comme origine choisie au hasard l'entree-gistement relatif à la huitième UL dans l'ensemble de données complet, puis en sélectionnant chaque quinzième enregistrement par la suite. L'ensemble de données résultant comporte les enregistrements de la NCS relatifs à 12,432 UL. Comme les UL figurant sur le fichier longitudinal initial sont ordonnées de telle façon que les unités appartenant à la même grappe sont regroupées, l'échantillon systématique de 1 sur 15 ne devrait pas comprendre plus d'une UL appartenant à la même grappe. En conséquence, le présent article ne traite pas du problème des corrélations entre les UL faisant partie d'une même grappe.

2.3 Flux dans les catégories d'actes criminels déclarés

Nous avons utilisé les données hiérarchiques longitudinales afin d'établir, pour les années 1975, 1976, 1977 et 1978, des matrices récapitulatives illustrant les flux au sein des catégories d'actes criminels déclarés entre les premières et deuxième interviews réalisées auprès d'un ménage au cours d'une même année. Il convient de noter que, les interviews de la NCS s'échelonnant sur toute l'année, la première interview peut avoir lieu à n'importe quel moment entre les mois de janvier et juin, tandis que la seconde peut prendre place entre les mois de juillet et décembre. Les actes criminels déclarés au cours de la première interview se sont produits entre les mois de juillet et de mai précédents, tandis que les actes déclarés au cours de la seconde se sont produits entre les mois de janvier et novembre. Ainsi, la présente analyse porte uniquement sur les crimes déclarés d'une interview à l'autre. Elle ne saurait, par exemple, aborder les questions relatives aux variations dans les déclarations des répondants à diverses périodes de l'année, si ce n'est de façon très générale.

Il convient de noter qu'un échantillon de ménages de la NCS a participé, pendant la période de collecte des données, à une expérience portant sur la période de référence. Comme les occupants des UL ayant participé à l'expérience devaient déclarer les actes criminels dont ils avaient été victimes pour des périodes de référence autres que la période habituelle de six mois, ces UL n'ont pas été retenues aux fins de cette analyse.

En vue de la présente analyse, chaque ménage interviewé au moins une fois au cours d'une année donnée a été classé en fonction des codes de déclaration et de réponse qui lui ont été attribués au moment des deux interviews. Les actes criminels peuvent avoir été déclarés par n'importe quel membre du ménage et constituer des crimes contre la personne ou contre le ménage. Les analyses dont il est fait état à la section 4 portent sur deux jeux de matrices illustrant les catégories d'actes criminels déclarés. Ces matrices sont présentées à l'annexe I.

NCS portant sur la totalité de la période de trois ans et demi que pour aussi peu que 25% de l'ensemble des personnes interviewées. De plus, l'occurrence de la non-réponse n'est pas aléatoire, c'est-à-dire qu'elle varie selon que le répondant a été ou non victime d'actes criminels voir, par exemple, Saphire (1984).

Dans cet article, nous élargissons les modèles d'estimation des flux bruts tenant compte de la non-réponse non aléatoire élaborés par Stasny (1986). En particulier, les modèles que nous allons étudier prévoient des matrices de flux symétriques dans les catégories d'actes criminels ainsi qu'une non-réponse parfaitement aléatoire, une non-réponse non aléatoire dont on n'a pas à tenir compte, ou une non-réponse dont il faut tenir compte.

On trouve à la section 2 de l'article une brève description de la NCS et des données longitudinales recueillies dans le cadre de l'enquête. La section 3 expose la forme générale des modèles décrivant la symétrie des matrices de flux bruts avec données manquantes et présente les méthodes itératives permettant d'obtenir les estimateurs du maximum de vraisemblance (EMV) pour les paramètres des modèles. La section 4 porte sur l'ajustement des modèles aux données de la NCS. Enfin, on trouve à la section 5 un énoncé des conclusions qu'on peut tirer de la présente étude ainsi qu'un exposé de certaines pistes de recherche.

2. LA NATIONAL CRIME SURVEY ET SES DONNÉES

2.1 Plan de sondage

L'échantillon de la NCS est un échantillon d'UL à plusieurs degrés stratifié. Mise en oeuvre en juillet 1972 par la Law Enforcement Assistance Administration, l'enquête est réalisée par le Bureau of Justice Statistics depuis décembre 1979. La population cible de la NCS comprend tous les membres âgés de 12 ans ou plus de la population civile occupant des unités de logement, à l'exception des pensionnaires d'établissements institutionnels. L'enquête permet de recueillir des données sur les crimes contre la personne et contre les ménages dont sont victimes les occupants des UL sélectionnées. La NCS porte sur les infractions et les tentatives d'infraction suivantes: les voies de fait, le vol d'automobile ou de véhicule à moteur, le vol avec effraction, le vol simple, le viol et le vol qualifié. Figurent au nombre des actes criminels ne faisant pas partie du champ d'observation de l'enquête: l'enlèvement, le meurtre, le vol à l'étalage et les infractions perpétrées dans les établissements commerciaux.

La NCS est une enquête par panel avec renouvellement pour laquelle chaque UL sélectionnée fait partie de l'échantillon pendant trois ans et demi. Comme les interviews sont réalisées à des intervalles de six mois, chaque occupant des UL sélectionnées peut participer à un total de sept interviews. Toutefois, la première interview réalisée à chaque UL constitue une interview repère (n.d.t. Les données recueillies dans le cadre de cette interview permettent de déterminer si les données obtenues à l'occasion de l'interview suivante portent réellement sur la période de référence) et n'est pas utilisée aux fins de l'établissement d'estimations. Bien qu'un intervalle de six mois s'écoule entre les interviews réalisées à chaque UL, les interviews de la NCS s'échelonnent sur toute l'année. En effet, afin d'assurer une utilisation efficace des intervieweurs, des interviews sont réalisées chaque mois auprès d'un sixième des UL formant l'échantillon. Comme c'est l'UL qui constitue l'unité d'échantillonnage de la NCS, rien n'est fait pour suivre les personnes qui quittent l'UL au cours de la période de trois ans et demi. On a plutôt choisi d'inclure dans l'univers de l'enquête les nouvelles personnes qui emménagent dans l'UL. Chaque groupe distinct de personnes occupant une UL pendant que cette dernière fait partie de l'échantillon de la NCS est considéré comme un ménage distinct.

Les interviews de la NCS portent sur toutes les personnes de 12 ans ou plus qui occupent l'UL sélectionnée au moment de l'interview. L'interview a pour objet d'interroger les répondants sur les crimes contre la personne ou contre le ménage dont ils ont été victimes au cours des six mois antérieurs. L'interview consiste à poser à chaque ménage répondant une série de six questions de sélection visant à recueillir des renseignements sur les crimes commis contre

Symétrie des flux, en tenant compte de la non-réponse, dans les catégories d'actes criminels déclarés par les victimes

ELIZABETH A. STASNY¹

RÉSUMÉ

La United States National Crime Survey (enquête nationale sur la criminalité aux États-Unis) est une enquête ménage de grande envergure ayant pour objet l'établissement d'estimations des pourcentages de personnes ou de ménages victimes d'actes criminels. L'enquête est une enquête par panel avec renouvellement et, selon le plan de sondage, les unités de logement sélectionnées font partie de l'échantillon pendant trois ans et demi, les occupants de ces unités de logement étant interviewés tous les six mois. Comme aussi peu que 25% des personnes interviewées dans le cadre de l'enquête participent à tous les cycles d'interview au cours de la période de trois ans et demi, la non-réponse constitue un des graves problèmes entachant les données longitudinales recueillies à l'aide de la National Crime Survey. De plus, l'occurrence de la non-réponse n'est pas aléatoire, c'est-à-dire qu'elle varie selon que le répondant a été ou non victime d'actes criminels. Le présent article porte sur des modèles d'estimation des flux bruts dans les catégories relatives à deux modes de classement des actes criminels déclarés: selon le nombre d'actes criminels et selon la gravité de l'acte criminel. Dans ces modèles, il y a des mécanismes de non-réponse aléatoire ou non aléatoire et les probabilités se rattachant aux flux bruts peuvent ne pas être assumées à des contraintes ou être symétriques. Les modèles sont ajustés aux données recueillies dans le cadre de la National Crime Survey à l'aide d'estimateurs du maximum de vraisemblance.

MOTS CLÉS: Données qualitatives; non-réponse dont on n'a pas à tenir compte; enquête longitudinale; National Crime Survey; non-réponse dont il faut tenir compte.

1. INTRODUCTION

La NCS (United States National Crime Survey/enquête nationale sur la criminalité aux États-Unis) est une enquête-ménage de grande envergure réalisée par le U.S. Bureau of the Census pour le compte du Bureau of Justice Statistics. Les données de la NCS servent à établir des estimations trimestrielles des pourcentages de personnes et de ménages victimes d'actes criminels ainsi que des estimations annuelles des taux de criminalité. L'enquête utilise un panel d'unités de logement (UL) avec renouvellement et, selon le plan de sondage, les personnes occupant les UL sélectionnées sont interviewées jusqu'à un maximum de sept fois, à des intervalles de six mois.

Les personnes interviewées dans le cadre de la NCS sont interrogées au sujet des crimes contre la personne ou contre la propriété dont elles ont été victimes au cours des six mois antérieurs. Dans le présent article, nous commençons par étudier les actes criminels déclarés, d'une interview à l'autre, par les ménages (M) occupant les UL sélectionnées. Pour chaque ménage, ces déclarations sont étudiées sous deux aspects: selon le nombre d'actes criminels déclarés (zéro, un et deux ou plus) et selon le genre d'acte criminel déclaré (aucun crime, crime contre la propriété et crime contre la personne).

Comme on ne dispose pas de réponses pour tous les ménages d'un cycle d'interview à l'autre, il nous faut déterminer de quelle façon il convient de traiter les données manquantes. Les données longitudinales recueillies dans le cadre de la NCS soulèvent un grave problème de non-réponse. Ainsi, Fienberg (1980) a noté qu'on ne dispose d'enregistrements d'interview de la

¹ Elizabeth A. Stasny, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, E.-U.

- BAILLAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: Wiley, 1-24.
- BIDERMAN, A.D., et CANTOR, D. (1984). A longitudinal analysis of bounding, respondent conditioning, and mobility as sources of panel bias in the national crime survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 708-713.
- CANTOR, D. (1985). Operational and substantive differences in changing the NCS reference period. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.
- GIESEMAN, R. (1987). The consumer expenditure survey: Quality control by comparative analysis. *Monthly Labor Review*, mars, 8-14.
- JOHNSON, R.A., et WICHERN, D.W. (1988). *Applied Multivariate Statistical Analysis*. 2ième édition, Englewood Cliffs, New Jersey: Prentice Hall, 188-190.
- KALTON, G., KASPRZYK, D., et McMILLEN, D. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: Wiley, 249-270.
- LOFTUS, E. (1986). Survey remembering. *Proceedings of the Second Annual Research Conference, Bureau of the Census*, Washington, D.C., 193-207.
- MATHIOWETZ, N.A. (1985). The problem of omissions and telescoping error: New evidence from a study of unemployment. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 482-487.
- MINGAY, D.J. (1987). Report on the consumer expenditure survey. *Questionnaire Design: Report on the 1987 BLS Advisory Conference*, (Eds. J. Bienias, C. Diplo, et M. Palmisano). Bureau of Labor Statistics, Washington, D.C., 129-138.
- MURPHY, L.R., et COWAN, C.D. (1976). Effects of bounding on telescoping in the National Crime Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 633-638.
- NETER, J., et WAKSBERG, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- NETER, J., et WAKSBERG, J. (1965). Response errors in collection of experimental data by household interviews: An experimental study. Technical Report No. 11, Bureau of the Census, Washington, D.C.
- SILBERSTEIN, A.R. (1988). Selected first-interview effects in the Consumer Expenditure Interview Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 485-490.
- SILBERSTEIN, A.R., et JACOBS, C.A. (1989). Symptoms of repeated interview effects in the Consumer Expenditure Interview Survey. *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, et M.P. Singh). New York: Wiley, 289-303.
- SILBERSTEIN, A.R. (1989). Recall effects in the U.S. Consumer Expenditure Interview Survey. *Journal of Official Statistics*, 2, 125-142.
- SUDMAN, S., et BRADBURN N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.
- U.S. BUREAU OF LABOR STATISTICS (1986). *Consumer Expenditure Survey: Interview Survey, 1984*. Washington, D.C., Bulletin No. 2267.

BIBLIOGRAPHIE

ANNEXE

(1) Description de certaines catégories de dépenses

VÊTEMENTS

Vêtements divers et vêtements combinés: vêtements de nuit, vêtements de détente, accessoires, uniformes et vêtements pour enfants de moins de 2 ans.

Autres vêtements et accessoires, services vestimentaires: montres, bijoux, matériel de couture pour la confection, services de réparation et de retouche, et location et entreposage de vêtements.

ARTICLES ET ACCESSOIRES D'AMEUBLEMENT

Autres appareils: petits appareils électriques pour la cuisine et les soins personnels.

Gros articles ménagers et matériel de détente: tondeuses à gazon, climatiseurs d'appareil, téléviseurs, chaînes stéréo et bicyclettes.

Autres articles ménagers et autre matériel de détente: postes de radio, magnétophones, outils, calculatrices, matériel de camping ou de sport, et matériel de divertissement pour enfants.

(2) Estimation de l'effet de télescopage

(d'après *Neter et Waksberg (1965), p. 33-37*).

Pour chaque catégorie de dépenses

Soit: x_U = moyenne d'échantillon pour période de référence d'un mois non délimitée; x_B = moyenne d'échantillon pour période de référence d'un mois délimitée (non observée directement dans la CE Interview Survey); x_2, x_3 = moyenne d'échantillon d'un mois pour les cycles 2 et 3 respectivement, calculées au moyen des premier et second mois de référence.

Définissons: l'effet de télescopage β , dans l'hypothèse où il n'y a pas d'effet de conditionnement,

(1)

$$\beta = (E x_U / E x_B) - 1;$$

l'effet de conditionnement α entre deux cycles successifs

(2)

$$\alpha = 1 - (E x_{t+1} / E x_t).$$

Alors, en supposant que le conditionnement vient s'ajouter au télescopage,

(3)

$$\beta_C = (E x_U / E x_B) (1 - \alpha) - 1$$

est l'effet de télescopage dans l'hypothèse où il y a un effet de conditionnement.

Étant donné l'effet de conditionnement estimé entre le second et le troisième cycle, $a = 1 - (x_3/x_2)$, la moyenne estimée pour une période de référence d'un mois délimitée est:

(4)

$$\begin{aligned} x_B &= (x_2 + x_3)/2 \\ &= (x_2 + x_2(1 - a))/2 \\ &= x_2(1 - a/2). \end{aligned}$$

Si nous supposons un effet de conditionnement constant et que nous utilisons les équations (3) et (4), l'effet de télescopage estimé, b_C , est:

(5)

$$b_C = (x_U/x_B) (1 - a) (1 - a/2) - 1.$$

4. CONCLUSIONS

Cet article a permis d'analyser l'effet de télescopage qui peut se produire dans les interviews qui n'ont pas de période de référence délimitée. Cet effet peut être considérable, surtout pour ce qui a trait aux événements qui laissent un souvenir plus profond. Ainsi, selon des données de la Consumer Expenditure Interview Survey aux E.-U., les estimés des montants consacrés à l'achat de biens durables à l'occasion d'une interview avec période de référence d'un mois non délimitée peuvent être de 30 à 50% supérieurs aux montants réels des dépenses. Cet écart est normalement moindre en ce qui concerne l'achat de biens non durables ou semi-durables. Ces observations viennent confirmer les résultats d'autres études faites sur la question. Notre étude a montré que l'effet de télescopage externe est beaucoup plus grand que l'effet de télescopage interne pour la période de référence de trois mois des cycles subséquents. De plus, elle a montré que les dépenses moyennes pour le premier cycle de l'enquête par panel étudiée étaient plus élevées que les dépenses moyennes pour l'ensemble des cycles subséquents, même après avoir déduit l'effet de télescopage estimé. Comme le premier cycle de cette enquête a une période de référence d'un mois, nous pouvons en conclure qu'en réduisant la période de référence, on peut s'attendre à des estimations de bien meilleure qualité. On pourrait penser à une amélioration d'au moins 10% dans le cas des dépenses qui se répètent souvent; toutefois, ce pourcentage deviendrait négligeable pour les dépenses de plus grande envergure qui ne se répètent à peu près pas.

Bien que la période de référence d'un mois est ce qui explique principalement le niveau plus élevé des estimations, il y a d'autres facteurs à considérer. Les effets de conditionnement, qu'on a supposés constants dans cette étude, peuvent varier avec les cycles. Les estimés basés sur une période de référence d'un mois seraient encore plus élevés si on supposait des effets de conditionnement plus importants entre le premier et le deuxième cycle. Afin de mieux comprendre le conditionnement des panels, on devrait faire des recherches sur les aspects cognitifs de l'interview tels la coopération et l'implication des répondants, et l'attitude des intervieweurs face à la collecte des données. Dans le même ordre d'idées, on devrait aussi se pencher sur les différences d'effets entre les catégories de dépenses. Des études sur le terrain et en laboratoire portant sur ces aspects de la collecte des données pourraient améliorer la méthodologie des enquêtes par panel.

REMERCIEMENTS

L'auteure tient à remercier les arbitres ainsi que Stuart Scott et Sylvia Leaver pour leurs commentaires.

Tableau 5
 Comparaison sommaire du premier cycle et des cycles subséquents
 Dépenses annuelles moyennes (erreurs types)

Cycle 1 (déduction faite de l'effet de télescopage)	Avec effet de condi- tionnement	Cycle 1 Cycles 2 à 5 Tous les mois de la période de référence (a)	Cycle 1 Cycles 2 à 5 Tous les mois de la période de référence	Sans effet de condi- tionnement	ARTICLES ET ACCESSOIRES					VÊTEMENTS					
					(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	
						\$1,663 (59.6)	\$1,182 (61.7)	\$1,452 (71.0)	\$1,295 (66.2)	\$1,370 (n.a.)					
						\$1,972 (85.0)	\$1,179 (59.7)	\$1,327 (73.1)	\$1,209 (61.5)	\$1,235 (n.a.)					

(a) Ces moyennes diffèrent des estimations publiées pour 1984 à cause du sous-ensemble particulier de répondants et de l'absence des coefficients de pondération finals.

les cycles subséquents sont présentés comme une forme de suivi pour le premier cycle, dans la CE Interview Survey ils portent sur une période trois fois plus longue que celle du premier cycle et les répondants doivent fournir des renseignements détaillés sur leur revenu. Ce fardeau de réponse supplémentaire ainsi que le rythme accru de l'interview, qui en découle, influent de façon négative sur les dépenses déclarées, même pour le premier mois de la période de référence de ces cycles. Par ailleurs, on peut recourir à un plus grand nombre de pièces justificatives (ex.: carnets de chèques, factures) dans les cycles subséquents, ce qui réduit les risques de télescopage. Bien que propice au télescopage externe, l'interview réalisée dans le premier cycle est plus simple que les interviews subséquentes, surtout pour ce qui a trait aux catégories de dépenses susceptibles de varier beaucoup selon la durée de la période de référence et la taille du ménage (par ex.: vêtements). L'importance relative de ces facteurs devrait être analysée sur le terrain et dans les laboratoires de psychologie cognitive.

Nous avons établi des estimations indépendantes des dépenses moyennes pour le premier cycle, déduction faite de l'effet de télescopage, en nous servant des deux séries d'estimations du tableau 4. Nous avons calculé ces dépenses moyennes (\bar{X}_{B1}) en divisant les dépenses moyennes pour la période de référence non délimitée par l'effet de télescopage estimé:

$$X_{B1} = \bar{X}_U / (1 + b_c).$$

Les résultats pertinents sont résumés dans le tableau 5 selon chaque groupe de biens. Dans les deux cas, les dépenses moyennes estimées pour le premier cycle (déduction faite de l'effet de télescopage) sont supérieures aux dépenses moyennes estimées pour l'ensemble de la période de référence des cycles subséquents (voir colonne 2). En ce qui concerne les vêtements, l'écart est de 10% lorsqu'il n'y a pas d'effet de conditionnement et de 16% dans le cas contraire. Pour ce qui a trait aux articles et accessoires d'ameublement, l'écart est moins prononcé: 3% lorsqu'il n'y a pas d'effet de conditionnement et 5% dans le cas contraire. On peut expliquer ces écarts par le fait que le premier cycle d'interview est caractérisé par une période de référence plus courte et un fardeau de réponses moins exigeant. Les différences de moyennes entre les deux groupes de biens et les résultats observés pour certaines catégories de dépenses impliquent que la qualité des renseignements fournis tend à s'accroître dans ce cas pour les articles peu chers, tandis que pour les articles d'une catégorie de prix supérieure, ce gain devient négligeable.

Tableau 4
Effet de télécopage estimé, selon la catégorie de dépenses

Effet de condition- nement <i>a</i>	Effet de télécopage <i>b_c</i>			
	<i>a</i> = 0	<i>a</i> > 0	<i>s</i>	<i>s</i>
	(1)	(2)	(3)	(4)
VÊTEMENTS:				
Manteaux, vestes, fourrure, tailleurs	28.4	7.0	-	-
Pantalons, jeans	46.2	14.2	-	-
Chemisiers, blouses, corsages	30.3	8.6	12.3	11.8
Chandails, robes, jupes	27.7	7.8	17.6	16.7
Vêtements de base, bas et chaussettes	28.3	5.9	8.7	15.0
Vêtements divers et vêtements combinés	22.2	6.9	7.2	12.7
Chaussures	5.2	9.5	-	-
Autres vêtements et accessoires, services vestimentaires	18.1	7.1	-	-
	54.9	35.8	-	-
ARTICLES ET ACCESSOIRES				
D'AMEUBLEMENT:	63.1	8.9	-	-
Gros appareils ménagers	95.4	30.7	-	-
Autres appareils	76.4	16.1	36.0	19.7
Gros articles ménagers et matériel	113.3	25.2	-	-
de détente	38.7	13.1	36.5	33.7
Autres articles ménagers et autre	26.2	8.9	-	-
matériel de détente	26.2	8.9	-	-
Accessoires d'ameublement -	15.6	14.5	-	-
réparations et services	15.6	14.5	-	-
Vaisselle, articles de décoration,	45.4	14.4	-	-
linge de maison	89.4	38.0	66.8	68.7
Couvre-planchers et cache-fenêtres	45.4	14.4	-	-

a Effet lié au nombre de mois passés dans l'échantillon, ou effet de conditionnement, si positif.
s Erreur type de la différence en pourcentage.

3.2 Effets du premier cycle

Les différences entre les réponses données dans le premier cycle d'interviews et celles données dans les cycles subséquents reflètent de nombreux aspects cognitifs des interviews de panel. Dans cette sous-section, nous allons examiner quelques-uns des facteurs en jeu et faire une analyse préliminaire des effets nets. Dans la mesure où les répondants participent à tous les cycles d'une enquête, il se crée progressivement des liens entre le répondant et l'intervieweur et chacun voit plus clairement ce que l'autre attend de lui. Toutefois, de nombreuses conditions peuvent changer dans l'interview. Par exemple, tandis que dans certaines enquêtes par panel

Nous avons posé un certain nombre d'hypothèses afin d'estimer l'effet de télécopage à l'aide des données d'enquête. Nous savons que les dépenses moyennes pour une période délimitée d'un mois servent à la comparaison avec le premier cycle d'interviews; or, on ne peut calculer directement ces dépenses à l'aide des données de la période de référence de trois mois. Diviser par trois les dépenses moyennes pour la période délimitée de trois mois n'est pas une solution, compte tenu de la baisse du taux de déclaration observée pour le troisième mois de référence de la CB Interview Survey. Par conséquent, nous avons plutôt utilisé les données des premier et second mois de référence pour estimer les dépenses moyennes pour une période délimitée d'un mois, en supposant que le biais de rappel dans le premier mois de référence se rattache à la plupart des événements qui sont situés à tort dans le premier mois de référence et que en réalité au second mois de référence. La méthode d'estimation utilisée est une adaptation du modèle qu'ont élaboré Netter et Waksberg dans leur analyse de l'étude expérimentale des dépenses au titre de la rénovation domiciliaire, réalisée en 1960 (Netter et Waksberg 1964 et 1965). Le modèle suppose que les effets de télécopage et de conditionnement sont multipli-catifs et que le conditionnement est lié au nombre de mois passés dans l'échantillon. Comme l'effet de conditionnement est déterminé à partir des relations observées entre le second et le troisième cycle, il nous faut deux termes pour estimer (2) suivant l'hypothèse du condition-nement. Par conséquent, nous pouvons estimer l'effet de télécopage au moyen de la formule suivante:

$$b_c = (x_U/x_B) (1 - a) (1 - a/2) - 1. \tag{4}$$

La façon de déterminer cette formule est décrite en annexe. Nous avons supposé que l'effet de conditionnement (a) était constant d'un cycle à l'autre, compte tenu du sous-ensemble particulier de répondants qui ont participé aux cinq cycles de l'enquête. (Le modèle de Netter et Waksberg supposait un effet de conditionnement plus élevé entre le premier et le second cycle.) L'effet lié au nombre de mois passés dans l'échantillon (ou effet de conditionnement) semble peu important dans la CB Interview Survey, s'il faut en croire une étude où l'on a comparé les réponses des cycles 2 à 5 (Silberstein et Jacobs 1989). Cela s'explique peut-être par le fait que la diminution graduelle du nombre de réponses est compensée par une améliora-tion de la qualité des déclarations, étant donné que les répondants connaissent de mieux en mieux le processus de déclaration. À l'aide de l'équation (4), nous avons pu calculer des estimations de l'effet de télécopage suivant deux hypothèses relatives au conditionnement: $a = 0$ (absence de l'effet de conditionnement) et $a > 0$ (effet de conditionnement égal à l'effet observé entre le second et le troisième cycle). Pour quatre catégories de dépenses du groupe vêtements et trois catégories d'articles et accessoires d'ameublement, on observe une diminution des dépenses moyennes déclarées entre le second et le troisième cycle; cette dimi-nution est représentée par des proportions positives dans la colonne 5 du tableau 4. Bien qu'elles ne soient pas significatives (à un seuil de .05), ces proportions sont également utilisées pour représenter l'effet de conditionnement entre le premier et le second cycle. On a considéré qu'une augmentation nette des dépenses déclarées entre ces deux cycles n'était pas réaliste. Les résultats nous donnent une idée de l'augmentation que subiraient les estimations s'il n'y avait pas de délimitation. Le tableau 4 contient des estimations de l'effet de télécopage (en pourcentage) dans le cas où il n'y a pas d'effet de conditionnement (col. 1) et dans le cas où il y en a (col. 3). Un effet de télécopage de plus de 40% est calculé pour les catégories "Manteaux, vestes, etc." et "Autres vêtements et accessoires, services vestimentaires" (cette dernière comprenant les montres et les bijoux); cependant, pour les autres catégories du groupe vêtements, l'effet de télécopage est beaucoup moins élevé. En ce qui concerne les articles et accessoires d'ameublement, l'effet de télécopage est élevé (63% en moyenne). L'effet de télécopage estime diminuer considérablement lorsqu'on tient compte de l'effet de condition-nement et diminuerait davantage si l'on supposait un effet de conditionnement plus grand entre le premier et le second cycle. Bien que ces estimations soient influencées par la variabilité

Conformément à ce qu'ont révélé les comparaisons précédentes, le test global est significatif et il y a beaucoup plus de répondants qui déclarent des dépenses de \$100 et plus dans le premier cycle; en revanche, la différence entre les taux de déclaration pour les divers cycles n'est pas significative dans le cas des dépenses moins importantes. Lorsqu'on examine les taux de déclaration pour les trois mois de référence des cycles 2 à 5, on s'aperçoit que les taux pour le premier mois se rapprochent plus des taux du premier cycle que de ceux des deux autres mois. On remarque de plus une diminution progressive du taux de déclaration, ce qui est peu surprenant; notons par ailleurs l'accroissement du pourcentage de répondants qui ne déclarent aucune dépense. Cette tendance ressort clairement dans chaque cycle d'interview d'un panel, comme en font foi l'article de Silberstein et Jacobs (1989) et celui de Silberstein (1989), et est probablement plus attribuable au biais de rappel qu'au télescopage. Lorsqu'on recalcule les taux de déclaration en tenant compte uniquement des répondants qui déclarent une dépense, on constate une plus grande similitude entre les trois mois de référence des cycles subséquents qu'entre l'un ou l'autre de ces trois mois et le premier cycle. (On peut calculer les taux à l'aide du tableau 3 en se servant du pourcentage de ceux ayant déclaré des dépenses comme base de calcul.) Ainsi, les taux de déclaration pour les articles ménagers de \$100 et plus sont de 53% dans le premier cycle et de 40, 41 et 40% respectivement pour les trois mois de référence des autres cycles. Pour ce qui a trait aux articles vestimentaires de \$100 et plus, les taux sont de 24% dans le premier cycle et de 19, 19 et 16% respectivement pour les trois mois de référence des autres cycles. Ces écarts seraient l'indice d'un télescopage externe dans l'interview n'ayant pas de période de référence délimitée.

3. ESTIMATION DE L'EFFET DE TÉLESCOPAGE
ET DES EFFETS DU PREMIER CYCLE
D'INTERVIEWS

3.1 Effet de télescopage

L'hypothèse de l'égalité des moyennes suppose que le processus de réponse dans le premier cycle d'interviews est identique à celui pour le premier mois de référence des cycles subséquents. Or, les données ne vérifient pas cette hypothèse puisque des écarts ont été observés, indiquant ainsi la possibilité d'un télescopage externe dans le premier cycle. Les résultats tendent à confirmer l'idée émise par Loftus (1986, p. 196) et selon laquelle le télescopage interne et le télescopage externe découleraient de mécanismes cognitifs différents. On peut définir de façon générale le télescopage externe (β) (sur une base mensuelle et en supposant qu'il n'y a pas de conditionnement de panel) comme le rapport entre les dépenses moyennes déclarées pour une période de référence d'un mois non délimitée (moyenne d'échantillon x_U) et les dépenses moyennes déclarées pour une période de référence d'un mois délimitée (moyenne d'échantillon x_B):

(2)
$$\beta = (Ex_U/Ex_B) - 1.$$

Cette expression est peut-être exagérée car l'effet de conditionnement contribue à réduire la valeur de la moyenne calculée pour des périodes délimitées. On observe habituellement une tendance à la baisse dans les réponses d'un panel parce que les répondants sont de moins en moins enclins à prêter leur concours à mesure que la période d'enquête s'écoule (Bailar 1989). On peut définir l'effet de conditionnement (α) entre deux cycles successifs par le rapport des deux réponses données (moyennes d'échantillon x_i et x_{i+1}):

(3)
$$\alpha = 1 - (Ex_{i+1}/Ex_i).$$

On peut interpréter ces différences de nombreuses façons. Par exemple, elles peuvent vouloir dire que les répondants déclarent des achats plus coûteux dans le premier cycle d'interviews ou qu'ils se rappellent leurs achats dans ce cycle comme des achats plus coûteux. On peut aussi penser que les réponses données portent sur une période plus longue qu'un mois lorsqu'il n'y a pas de période de référence délimitée, surtout si l'on s'agit d'une grosse dépense qu'il est facile de se rappeler. Dans le tableau 3, nous poussons plus loin la comparaison des cycles en considérant cette fois les trois mois de référence des cycles subséquents. Les résultats concordent avec ceux des tests antérieurs mais tendent à faire ressortir essentiellement l'effet de téléscopage. La comparaison se fait au moyen de taux de déclaration exprimés en fonction de la valeur monétaire des dépenses déclarées. On définit le taux de déclaration comme le pourcentage de répondants qui déclarent au moins une dépense d'une catégorie donnée. Notons que chaque dépense est habituellement inscrite sur le questionnaire, sauf s'il s'agit d'une dépense qui revient régulièrement au cours d'un mois pour un membre du ménage; dans un tel cas, on ne considère qu'un montant global pour le mois.

Tableau 3

Taux de déclaration mensuels selon le niveau de dépenses

Cycles 2 à 5 par mois de référence	Cycle 1		
	Premier	Deuxième	Troisième
	Pourcentage de répondants		
	(1)	(2)	(3)
	(4)		

VÊTEMENTS:

Aucune dépense (a)

Moins de \$10

De \$ 10 à \$ 40

De \$ 40 à \$100

\$100 et plus

Cycle 1 par rapport au premier mois de

référence des cycles 2 à 5

Valeur du test global: 29.1*

ARTICLES ET ACCESSOIRES

D'AMEUBLEMENT:

Aucune dépense (a)

Moins de \$10

De \$ 10 à \$ 40

De \$ 40 à \$100

De \$100 à \$400

\$400 et plus

Cycle 1 par rapport au premier mois de

référence des cycles 2 à 5

Valeur du test global: 17.0*

(a) Catégorie incluse dans le test global.
* Significatif à un seuil ($\alpha = .05$).

2.2 Résultats des tests

Dans le tableau 1, nous comparons des moyennes dans leur forme originale, c'est-à-dire sans la transformation logarithmique utilisée dans les tests statistiques. Le premier cycle pré-sente des moyennes plus élevées dans presque tous les cas et le test global est significatif. Les tests appliqués à chaque catégorie de dépenses révèlent des différences significatives unique-ment dans le cas des postes de dépense majeurs, comme les manteaux et les vestes, du groupe vêtements, et les appareils et les meubles, du groupe articles et accessoires d'ameublement. Les catégories de dépenses pour lesquelles on a relevé des différences significatives sont plus fortement représentées dans le premier cycle que dans les autres cycles, ce qui n'est pas sur-prenant si on tient compte du fait que ces catégories représentent 19% des dépenses totales en vêtements et 72% des dépenses totales en articles et accessoires d'ameublement dans le premier cycle, comparativement à 16 et à 67% respectivement dans le premier mois de référence des autres cycles (voir colonnes 1 et 2 du tableau 2). En outre, les répondants déclarent un plus grand nombre de dépenses pour ces catégories dans le premier cycle (voir colonnes 3 et 4 du tableau 2). Enfin, la valeur monétaire moyenne des dépenses déclarées dans le premier cycle est généralement différente de celle des dépenses déclarées dans les autres cycles en ce qui a trait aux articles chers (par ex. : gros appareils ménagers) mais cette différence s'estompe dans le cas d'articles d'une catégorie de prix inférieure (voir colonnes 5 et 6 du tableau 2).

Tableau 2

Comparaison entre le premier cycle et le premier mois de référence des cycles subséquents

Pourcentage des dépenses totales		Pourcentage du nombre total de dépenses		Valeur monétaire moyenne des dépenses	
1 Cycle	1 à 5 Cycles	1 Cycle	1 à 5 Cycles	1 Cycle	1 à 5 Cycles
(1)	(2)	(3)	(4)	(5)	(6)
100.0	100.0	100.0	100.0	\$ 35	\$ 33
19.2	15.7	9.3	8.6	71	59
Pantalon, jeans	10.7	10.8	10.6	9.8	36
Chemisiers, blouses, corsets	10.0	10.4	12.0	12.2	31
Chandails, robes, jupes	14.3	14.0	13.0	12.4	38
Vêtements de base, bas et chaussettes	5.2	5.6	16.8	16.7	11
Vêtements divers et vêtements combinés	15.5	18.2	15.4	16.4	36
Chaussures	11.7	13.1	12.8	13.6	33
Autres vêtements et accessoires,	13.5	12.2	10.1	10.4	45
services vestimentaires					40
ARTICLES ET ACCESSOIRES D'AMEUBLEMENT:					
100.0	100.0	100.0	100.0	\$123	\$ 92
Gros appareils ménagers	11.4	9.6	4.2	3.4	370
Autres appareils	2.3	2.2	9.2	7.1	29
Meubles	28.3	19.9	8.9	7.5	385
Gros articles ménagers et matériel de détente	19.7	21.8	8.8	7.6	262
Autres articles ménagers et autre matériel de détente	10.7	13.4	22.7	22.8	58
Accessoires d'ameublement - réparations et services	4.7	6.6	8.4	9.5	67
Vaisselle, articles de décoration, linge de maison	12.9	16.8	33.1	37.5	46
Couvre-planchers et cache-fenêtres	10.0	9.8	4.6	4.5	294
					172
					39

Pour pouvoir faire des comparaisons par catégorie de dépenses, nous avons construit des intervalles de confiance simultanés au moyen de la méthode de Bonferroni (Johnson et Wichern 1988), avec comme percentile $t_n^*(.05/2p)$. Nous avons calculé les dépenses moyennes en soumettant à une transformation logarithmique les dépenses déclarées par chaque répondant pour le premier mois de référence. Les poids d'échantillonnage comprenaient des facteurs de correction pour la non-réponse et le sous-échantillonnage mais non les coefficients de pondération finals pour totaux de contrôle de la population, qui n'existaient pas pour le premier cycle d'interviews. Il convient de souligner que le calcul de coefficients de pondération pour le premier cycle ont été calculés uniquement pour cette analyse et que ces coefficients ne sont pas nécessaires pour l'estimation.

Nous avons groupé les données des cycles 2 à 5 puisqu'il y avait très peu de différence entre ces cycles. Nous avons retenu les réponses fournies par les personnes qui ont participé aux cinq cycles d'interviews (3200 répondants) de manière à assurer la comparabilité des cycles et la délimitation de la période de référence des cycles 2 à 5. Dans la CE Interview Survey, les nouveaux membres de panel (par ex.: les nouveaux occupants d'un logement échantillonné) et les répondants qui ne participent pas à tous les cycles de l'enquête sont soumis à une interview sans période de référence délimitée. En 1984, 89% des interviews réalisées dans les cycles 2 à 5 portaient sur une période délimitée et 11% portaient sur une période non délimitée (8% à cause de répondants nouveaux et 3% à cause d'une non-participation antérieure) (Silberstein 1988). Comme le soulignent Biderman et Cantor (1984), les réponses données dans des interviews sans période de référence délimitée influent sur les estimations, mais nous n'insisterons pas ici sur ce point.

Tableau 1

Différence des dépenses moyennes (en pourcentage)		
Premier cycle par rapport au premier mois de référence des cycles 2 à 5		
Différence (en %)		(a)
		s
VÊTEMENTS: (b)		
Manteaux, vestes, fourrure, tailleurs	39.6*	12.9
Pantalons, jeans	13.6	9.5
Chemisiers, blouses, corsages	9.7	5.6
Chandails, robes, jupes	16.4	4.7
Vêtements de base, bas et chaussettes	6.9	5.4
Vêtements divers et vêtements combinés	-2.5	7.3
Chaussures	2.1	6.1
Autres vêtements et accessoires, services vestimentaires	27.4	25.4
Valeur du test global: 4.16*		
ARTICLES ET ACCESSOIRES D'AMEUBLEMENT: (b)		
Gros appareils ménagers	76.1*	27.5
Autres appareils	56.3*	17.0
Meubles	111.0*	24.8
Gros articles ménagers et matériel de détente	34.2*	16.0
Autres articles ménagers et autre matériel de détente	19.1*	7.1
Accessoires d'ameublement - réparations et services	7.0	14.6
Vaisselle, articles de décoration, linge de maison	14.0	16.0
Couvre-planchers et cache-fenêtres	52.5	24.3
Valeur du test global: 13.86*		

(a) Une valeur positive indique que la moyenne pour le premier cycle est supérieure. La base de calcul du pourcentage est la moyenne pour le premier mois de référence des cycles 2 à 5.

(b) Le total pour le groupe n'est pas considéré dans le test global.

s Erreur type de la différence en pourcentage.

* Significatif à un seuil de 5% ($\alpha = .05$).

La toute première étape de l'analyse consiste à vérifier si les données d'enquête ne renfermeraient pas des signes de télescopage externe. Lorsque le premier cycle d'interviews produit des estimations plus élevées que prévu, on peut penser qu'il y a eu télescopage. Les interviews qui portent sur une période non délimitée produisent normalement des estimations plus élevées que les interviews ayant une période de référence délimitée, comme le révèlent plusieurs études où l'on compare les réponses fournies dans ces deux types d'interviews (Neter et Wakseberg 1964 et 1965; Murphy et Cowan 1976; Cantor 1985). On peut aussi penser qu'il y a eu télescopage lorsqu'on constate des différences d'effet entre les diverses catégories de dépenses. Le biais de rappel et le télescopage sont deux facteurs importants qui font que tous les répondants ne se rappellent ni ne déclarent les événements de la même façon. La relation entre ces deux facteurs permet de croire que le souvenir des dépenses d'importance secondaire tend à s'effacer rapidement alors que le souvenir des dépenses plus importantes, qui persiste plus longtemps, est souvent télescopé.

Le télescopage peut aussi se produire dans des interviews avec période de référence délimitée; dans un tel cas, les événements sont situés à une date postérieure à celle où ils se sont vraiment produits mais cela, dans les limites de la période de référence (c'est ce qu'on appelle le télescopage interne). Bien que ce genre de télescopage ne modifie en rien les estimations globales, il influe sur leur répartition à l'intérieur de la période de référence de trois mois. Pour les besoins de notre analyse, nous avons choisi les dépenses relatives aux vêtements et aux articles et accessoires d'ameublement à cause de l'utilité des caractéristiques de ces dépenses. Primo, il existe divers ordres de grandeur pour ces dépenses et les biens correspondants sont groupés en conséquence. Secundo, le degré de sous-déclaration n'est pas le même pour toutes les catégories de dépenses. De nombreuses estimations relatives aux vêtements sont 40% moins élevées que les estimations des Comptes nationaux (CN) et plusieurs estimations relatives aux articles et accessoires d'ameublement sont également inférieures aux estimations correspondantes des CN. Par ailleurs, les estimations de dépenses pour les meubles et certaines catégories d'accessoires ne sont que de 7% inférieures aux estimations indépendantes (Gieseeman 1987, p. 11), et les estimations relativement plus élevées dans le premier cycle d'interviews peuvent être vues comme la conséquence d'un télescopage externe plutôt que de la sous-déclaration. Notre analyse visait à vérifier l'hypothèse que les dépenses déclarées pour le premier mois de référence des cycles avec délimitation, c'est-à-dire le mois précédant l'interview, sont comparables aux dépenses déclarées pour le mois correspondant du premier cycle. Nous avons donc testé cette hypothèse pour huit catégories de dépenses dans chaque groupe (vêtements et articles et accessoires d'ameublement) en nous servant de la statistique T^2 d'Hotelling. Étant donné deux vecteurs de moyennes construits selon un plan avec répétition, nous avons testé l'hypothèse $H_0: C\mu = 0$ (égalité des moyennes) par rapport à $H_1: C\mu \neq 0$ au moyen d'un test bilatéral à un seuil de signification de .05. H_0 était rejetée si:

$$(1) \quad [(C\bar{x})'(CS C')^{-1} C\bar{x}] / [np/(n - p - 1)] > F_{p, n-p-1}(.05),$$

où \bar{x} est un vecteur de moyennes d'échantillon pour chaque groupe de biens (répartis suivant la classification présentée dans les tableaux), S est la matrice des covariances calculée au moyen de la méthode de la répétition compensée ($n = 20$ échantillons répétés), C est la matrice des contrastes ci-dessous et p est le nombre de contrastes dans C .

$$C^{(p \times 2p)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \end{bmatrix}.$$

Expenditure (CE) Interview Survey des Etats-Unis de 1984, nous analysons les effets que peuvent avoir sur les estimations des dépenses des ménages le télescopage, le conditionnement et la durée de la période de référence. La CE Interview Survey est une des deux enquêtes qui visent à recueillir des données nationales sur les dépenses des ménages, l'autre étant la Diary Survey. L'enquête est réalisée par le Census Bureau pour le compte du Bureau of Labor Statistics. Le premier cycle de la CE Interview Survey permet à l'intervieweur de s'assurer la collaboration des répondants, de recueillir des données de base sur les biens des ménages et de délimiter la période de référence du second cycle. Le premier cycle d'interviews est suivi de quatre autres cycles à trois mois d'intervalle les uns des autres; à chaque occasion, on recueille des données pour les trois mois précédant le jour de l'interview. La méthode de délimitation est comme suit. Les dépenses déclarées pour le mois où a lieu l'interview (ou "mois courant") sont transcrites ultérieurement sur le questionnaire du cycle suivant; l'intervieweur peut se servir de ces données pour vérifier si des renseignements ne se répètent pas mais ne les communique pas au répondant. Les données recueillies durant le premier cycle ont trait aux dépenses pour le mois courant et pour un mois antérieur; les données pour le mois courant sont utilisées dans le second cycle tandis que celles pour un mois antérieur sont totalement exclues des estimations. Pour plus de détails sur les méthodes de collecte et d'estimation, voir le Bulletin de 1984 (U.S. Bureau of Labor Statistics, 1986; Silberstein et Jacobs (1989) font aussi une analyse de ces méthodes.

Les résultats soulignent la nécessité de délimiter la période de référence lorsqu'on recueille des données rétrospectives sans quoi l'effet de télescopage pourrait être appréciable. L'analyse révèle aussi que les réponses données dans le premier cycle d'interviews peuvent produire des estimations plus élevées que dans les cycles suivants, même après qu'on a éliminé l'effet de télescopage. Cela peut être une conséquence directe de la durée relativement courte de la période de référence du premier cycle de l'enquête, bien que ce ne soit pas nécessairement le seul facteur. Dans la section 2, nous décrivons l'analyse permettant d'identifier l'effet de télescopage. Ensuite, nous estimons cet effet de même que les effets du premier cycle d'interviews. Enfin, dans la section 4, nous donnons les conclusions de notre analyse.

2. L'EFFET DE TÉLESCOPAGE

2.1 Méthode d'analyse

Une façon d'identifier l'effet de télescopage, selon Kalton et coll. (1989, p. 257), est de vérifier si des éléments de réponse ne se répèteraient pas d'un cycle à l'autre. Cette méthode n'est pas nécessairement précise puisque le répondant pour un ménage donné peut changer d'un cycle à l'autre. Elle n'est pas non plus pratique car on ne dispose pas toujours d'enregistrements indépendants, outils indispensables au rapprochement de dates. Dans une enquête permanente, l'intervieweur peut ne pas enregistrer des réponses qui ont déjà été données dans un cycle antérieur même s'il les identifie comme telles durant l'interview, comme cela se produit dans la CE Interview Survey. Toutefois, on évalue le plus souvent l'effet de télescopage au niveau global en comparant, non sans précautions, les données obtenues sans période de référence précise à celles obtenues avec une période de référence délimitée. Il est recommandé de suivre plusieurs panels afin d'assurer la comparabilité des données d'une période à l'autre, puisque les réponses qui concernent des périodes délimitées viennent après celles qui ont trait à des périodes non délimitées et que, par conséquent, il ne s'agit pas du même intervalle de temps. Un autre facteur qu'il faut considérer dans l'exercice de comparaison est le conditionnement du panel, phénomène voulant que la qualité des déclarations change ou que le répondant modifie son attitude du fait qu'il appartient à un panel. Les hypothèses et la méthode d'estimation qui ont servi à cette étude sont décrites dans la section 3; pour l'instant, nous nous intéresserons aux premières étapes de l'analyse.

Effets du premier cycle d'interviews dans la Consumer Expenditure Interview Survey aux E.-U.

ADRIANA R. SILBERSTEIN¹

RÉSUMÉ

Les réponses des panels de la Consumer Expenditure Interview Survey des E.-U. sont comparées afin d'évaluer l'importance du téléscopage dans le premier cycle, non délimité, de cette enquête. Les résultats de l'analyse de certaines catégories de dépenses viennent appuyer les conclusions d'autres études selon lesquelles le téléscopage peut être appréciable dans des interviews avec période de référence non délimitée et varie selon la catégorie de dépenses. De plus, nous en venons à constater que les estimations tirées du premier cycle sont supérieures à celles tirées des cycles subséquents, même après avoir éliminé l'effet de téléscopage; par ailleurs, on peut attribuer en grande partie cet effet à la durée relativement courte de la période de référence du premier cycle de l'enquête.

MOTS CLÉS: Délimitation; téléscopage; biais de rappel; conditionnement.

1. INTRODUCTION

Dans les enquêtes rétrospectives, on demande aux répondants de se rappeler les détails d'événements qui se sont produits dans un intervalle de temps particulier, appelé période de référence, et il est parfois aussi difficile de situer correctement les événements dans le temps que de se les rappeler. La confusion de dates, ou "téléscopage", est généralement reconnue comme une source d'erreur dans les enquêtes, bien qu'elle n'ait pas fait souvent l'objet d'études approfondies (Neter et Waksberg, 1965). Les répondants ont tendance à rapporter des événements qui se sont produits en dehors de la période de référence (téléscopage externe); par exemple, ils peuvent situer un événement à une époque plus récente que celle où il s'est vraiment produit (téléscopage en aval). Des données que l'on a pu confirmer par des enregistrements indépendants montrent que le téléscopage se fait aussi bien en amont qu'en aval (Mathiowetz, 1985). Ce phénomène est peut-être attribuable au souci du répondant de bien s'acquitter de sa tâche. Lorsqu'il est dans l'incertitude, le répondant préfère en dire plus que pas assez (Sudman et Bradburn, 1974, p. 69). Globalement, le téléscopage se fera surtout en aval. La délimitation vise à bien définir le début et la fin de la période de référence de l'enquête de manière à éliminer les risques de téléscopage. Une façon efficace de fixer le début de la période de référence durant l'interview même est de comparer les événements qui sont rapportés à cette occasion avec ceux qui ont été rapportés dans une interview antérieure et de supprimer les éléments d'information qui se répètent. Pour marquer la fin de la période de référence, on la fait souvent coïncider avec le jour de l'interview. La délimitation est forcément impossible à réaliser dans le cas des enquêtes uniques ou lorsqu'il s'agit de questions qui ne sont posées qu'une seule fois ou du premier cycle d'une enquête par panel puisque dans tous ces cas, il n'existe pas de données antérieures qui puissent servir de repères. Il y a moyen d'atténuer ce problème en ayant recours à des techniques d'"ancrage" durant l'interview (par ex.: utilisation d'une ligne du temps

Cet article a pour but d'analyser les données fournies par des personnes qui participaient pour la première fois à une enquête par panel et d'établir une comparaison entre le premier cycle d'interviews et les cycles subséquents. En nous servant de données de la Consumer

¹ Adriana R. Silberstein, statisticienne, Office of Prices and Living Conditions, Statistical Methods Division, U.S. Bureau of Labor Statistics, Washington, D.C. 20212, E.-U.

L'ajustement s'est fait à l'aide d'un modèle linéaire logarithmique, qui a servi à prédire les valeurs de τ_5 . Il semble satisfaisant sauf pour ce qui a trait à la corrélation entre x_2 et x_3 , pour laquelle la valeur de R^2 est très faible dans les deux provinces (N.-E. et ONT).

4.4 Conclusions

L'estimation de coefficients de corrélation à l'aide de données d'enquête complexe est un problème délicat, non à cause de la difficulté des formules – en fait, les formules utilisées dans cette étude sont élémentaires – mais plutôt à cause des nombreuses contraintes auxquelles est soumise l'application de ces formules. Si nous n'avions pas posé les hypothèses de la section 3, il aurait été impossible d'estimer les corrélations de panel à l'aide du programme dont nous disposons. Par ailleurs, ces hypothèses devraient être compatibles avec les données réelles auxquelles sont appliquées les formules. Dans cette analyse, il semble que les hypothèses posées ne cadrent pas tout à fait avec les données réelles, vu la divergence des résultats obtenus à l'aide des équations (3) et (4) (tableau 3). Néanmoins, nous ne jugeons pas que les estimations sont aberrantes.

Les résultats de cette étude ont servi de façon concluante à comparer divers estimateurs composites pour l'EPA (Kumar et Lee 1983). Récemment, Binder et Dick (1990) ont proposé une méthode d'analyse des modèles saisonniers ARMMI qui tient compte des erreurs d'enquête. Ils ont appliqué cette méthode aux données de l'EPA en se servant des corrélations de panel estimées. Toutefois, si les résultats obtenus à l'aide de ces corrélations dépendent largement de la précision des estimations, ils doivent être interprétés avec prudence.

REMERCIEMENTS

L'auteur tient à remercier les arbitres anonymes ainsi que le rédacteur en chef, M.P. Singh, et la rédactrice adjointe L. Mach pour leurs précieux commentaires. Il remercie également S. Kumar et Y. Bélanger, de Statistique Canada, qui ont contribué, par leurs remarques constructives, à améliorer la version antérieure de cet article.

BIBLIOGRAPHIE

CHODHRY, G.H., et LEE, H. (1987). Estimation de la variance pour l'enquête sur la population active du Canada. *Techniques d'enquête*, 13, 157-172.

BINDER, D.A., et DICK, J.P. (1990). Méthode pour l'analyse des modèles ARMMI. *Techniques d'enquête*, ce numéro.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

KUMAR, S., et LEE, H. (1983). Évaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.

LEE, H. (1989a). Variance estimation methodology and general purpose variance estimation system for the Labour Force Survey. Methodology Working Paper Series, SSMD-89-022 E, Statistique Canada.

LEE, H. (1989b). Estimation of panel correlations for the Canadian Labour Force Survey. Methodology Working Paper Series, SSMD-89-023 E, Statistique Canada.

PLATEK, R., et SINGH, M.P. (1976). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue, Statistique Canada.

SINGH, M.P., DREW, J.D., GAMBINO, J., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue, Statistique Canada.

Les valeurs estimées des coefficients de corrélation γ figurent dans les tableaux 4A et 4B. Comme on pouvait s'y attendre, les valeurs de γ sont beaucoup moins élevées que celles de ρ mais elles reflètent, comme ρ , des différences de mobilité entre certains sous-groupes de la population active.

La tendance générale des γ est assez bizarre, surtout en ce qui concerne les estimations établies à l'aide des données de 1985-1987. Pour environ 25% des cas qui forment la tableau 4B – un cas équivaut à une ligne du tableau, on note une tendance à la hausse des valeurs estimées. En outre, dans la plupart de ces cas, la valeur de R^2 est peu élevée, ce qui indique la faible précision de l'ajustement obtenu à l'aide du modèle linéaire logarithmique. Cela ne signifie pas pour autant qu'il existe d'autres modèles qui pourraient nous permettre d'obtenir un meilleur ajustement. Tout ce que nous pouvons déduire de ces observations, c'est qu'aucune tendance claire ne ressort de ce tableau. Dans environ la moitié des cas pour lesquels on note une tendance à la baisse, la valeur de R^2 est supérieure à 0.5.

En ce qui concerne les estimations de γ établies à l'aide des données de 1980-1981, la situation est tout à fait différente. On note un seul cas où les valeurs estimées suivent une tendance à la hausse; de plus, la valeur de R^2 est supérieure à 0.5 pour la plupart des cas. De fait, pour ce qui a trait à la tendance, les estimations relatives à la N.-E. et à la C.-B. semblent plus conformes que celles pour l'ONT.

4.3 Valeurs estimées des coefficients de corrélation τ

Le tableau 5 renferme les valeurs estimées de τ , établies à l'aide des données de 1985-1987, pour toutes les combinaisons possibles de EMP 15-24 (x_3) et UNEMP 25 + (x_2), UNEMP 15-24 (x_3) et UNEMP 25 + (x_4). La corrélation entre x_1 et x_2 est essentiellement positive, tout comme celle entre x_3 et x_4 . Pour toutes les autres combinaisons, elle est essentiellement négative. En ce qui a trait au degré de corrélation, seuls les coefficients relatifs à (x_1, x_3) et à (x_2, x_4) sont sensiblement différents de zéro, les autres étant proches de zéro. Ces observations semblent confirmer ce que nous savons déjà du transfert de membres de la population active du même groupe d'âge entre le sous-groupe des personnes occupées et celui des personnes en chômage. Lorsque le nombre des personnes occupées augmente, celui des personnes en chômage diminue et vice-versa. La tendance est évidemment à la hausse dans ces cas.

Tableau 5
Valeurs estimées du coefficient de corrélation τ
 x_1 : EMP 15-24, x_2 : UNEMP 25 +, x_3 : UNEMP 15-24, x_4 : UNEMP 25 +,
(données de 1985-1987)

Province	Caractéristique	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
N.-E.	(x_1, x_2)	0.150	0.140	0.148	0.181	0.187	0.196
	(x_1, x_3)	-0.440	-0.275	-0.187	-0.135	-0.039	0.126
	(x_1, x_4)	-0.036	-0.040	-0.043	-0.015	0.024	0.022
	(x_2, x_3)	-0.029	-0.037	-0.078	-0.049	-0.016	-0.038
	(x_2, x_4)	-0.437	-0.374	-0.276	-0.182	-0.231	-0.094
	(x_3, x_4)	0.136	0.127	0.094	0.055	0.049	0.020
ONT	(x_1, x_2)	0.092	0.070	0.055	0.040	0.028	0.010
	(x_1, x_3)	-0.420	-0.267	-0.205	-0.161	-0.145	-0.010
	(x_1, x_4)	-0.065	-0.056	-0.053	-0.036	-0.028	-0.019
	(x_2, x_3)	-0.061	-0.054	-0.054	-0.042	-0.089	-0.074
	(x_2, x_4)	-0.392	-0.303	-0.230	-0.187	-0.181	-0.077
	(x_3, x_4)	0.058	0.043	0.022	0.013	0.022	0.001

Tableau 4A
Valeurs estimées du coefficient de corrélation γ
(données de 1980-1981)

Prov. Caractéristique	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}
N.-É. IN LF	0.288	0.263	0.265	0.250	0.236	0.233	0.211	0.199	0.193	0.167	0.164
EMP	0.262	0.219	0.228	0.226	0.219	0.239	0.210	0.200	0.188	0.161	0.172
EMP AG	0.351	0.308	0.283	0.237	0.205	0.190	0.141	0.113	0.063	0.021	0.007
EMP NON-AG	0.238	0.187	0.189	0.180	0.164	0.151	0.123	0.121	0.136	0.091	0.086
UNEMP	0.106	0.176	0.091	0.097	0.091	0.076	0.066	0.063	0.066	0.032	0.031
ONT IN LF	0.161	0.141	0.128	0.133	0.135	0.136	0.125	0.127	0.124	0.122	0.117
EMP	0.164	0.136	0.142	0.147	0.149	0.148	0.150	0.153	0.153	0.141	0.146
EMP AG	0.477	0.483	0.474	0.486	0.451	0.474	0.459	0.429	0.394	0.323	0.368
EMP NON-AG	0.184	0.150	0.147	0.157	0.163	0.167	0.166	0.169	0.174	0.156	0.165
UNEMP	0.141	0.074	0.076	0.063	0.080	0.051	0.045	0.060	0.077	0.136	0.074
C.-B. IN LF	0.177	0.137	0.117	0.119	0.119	0.112	0.101	0.112	0.094	0.066	0.070
EMP	0.211	0.146	0.133	0.107	0.101	0.083	0.050	0.068	0.058	-0.033	-0.015
EMP AG	0.380	0.311	0.301	0.272	0.241	0.216	0.198	0.170	0.122	0.078	0.071
EMP NON-AG	0.207	0.166	0.161	0.129	0.108	0.093	0.069	0.038	0.023	-0.004	-0.020
UNEMP	0.126	0.125	0.114	0.103	0.091	0.076	0.062	0.092	0.032	0.040	0.031

Tableau 4B
Valeurs estimées du coefficient de corrélation γ
(données de 1985-1987)

Prov. Caractéristique	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}
N.-É. IN LF	0.250	0.238	0.247	0.230	0.216	0.204	0.181	0.196	0.189	0.162	0.160
EMP	0.170	0.183	0.205	0.196	0.185	0.157	0.158	0.194	0.198	0.219	0.198
EMP AG	0.326	0.296	0.246	0.245	0.265	0.267	0.234	0.217	0.259	0.269	0.231
EMP NON-AG	0.146	0.168	0.199	0.201	0.178	0.153	0.152	0.189	0.199	0.216	0.201
UNEMP	0.233	0.267	0.241	0.211	0.206	0.168	0.171	0.176	0.157	0.187	0.147
EMP 15-24	0.107	0.127	0.140	0.133	0.112	0.105	0.099	0.107	0.090	0.074	0.082
EMP 25+	0.088	0.075	0.117	0.108	0.100	0.099	0.090	0.103	0.099	0.137	0.118
UNEMP 15-24	0.051	0.080	0.042	0.024	0.054	0.061	0.079	0.081	0.058	0.011	0.049
UNEMP 25+	0.155	0.129	0.177	0.171	0.148	0.159	0.158	0.127	0.102	0.134	0.124
ONT IN LF	0.162	0.138	0.141	0.134	0.132	0.135	0.127	0.116	0.111	0.103	0.101
EMP	0.114	0.122	0.121	0.122	0.117	0.124	0.119	0.108	0.110	0.112	0.111
EMP AG	0.508	0.518	0.553	0.561	0.571	0.569	0.582	0.617	0.668	0.650	0.672
EMP NON-AG	0.133	0.140	0.132	0.140	0.157	0.156	0.168	0.182	0.204	0.205	0.210
UNEMP	0.030	0.047	0.055	0.047	0.043	0.048	0.039	0.030	0.039	0.048	0.041
EMP 15-24	0.012	-0.006	0.018	0.031	0.017	0.023	0.011	0.016	0.016	0.044	0.029
EMP 25+	0.354	0.358	0.349	0.343	0.319	0.312	0.298	0.285	0.276	0.240	0.246
UNEMP 15-24	0.068	0.039	0.038	0.058	0.033	0.026	0.008	0.018	0.011	-0.002	-0.006
UNEMP 25+	0.052	0.054	0.033	0.017	0.034	0.033	0.026	0.018	0.021	0.044	0.022
C.-B. IN LF	0.103	0.095	0.113	0.103	0.090	0.090	0.091	0.083	0.078	0.030	0.055
EMP	0.125	0.100	0.112	0.111	0.116	0.135	0.123	0.121	0.118	0.095	0.114
EMP AG	0.394	0.443	0.426	0.401	0.396	0.400	0.401	0.381	0.347	0.334	0.345
EMP NON-AG	0.080	0.067	0.076	0.072	0.091	0.109	0.111	0.118	0.112	0.106	0.124
UNEMP	0.096	0.086	0.084	0.080	0.083	0.097	0.068	0.074	0.068	0.083	0.071

Comme prévu, les valeurs de ρ sont généralement élevées puisqu'elles mesurent la corrélation entre des estimations de panels communs. Elles sont le plus élevées pour EMP AG et le moins élevées pour UNEMP. C'est comme si la valeur de ρ indiquait le degré de mobilité d'un sous-groupe particulier de la population active. Par exemple, la valeur élevée de ρ pour la caractéristique EMP AG dénote une faible mobilité de la population active en agriculture, tandis que la faible valeur de ρ pour UNEMP indique une grande mobilité des personnes en chômage. La différence de mobilité de la population active entre les deux groupes d'âge étudiés est également très visible. Les plus jeunes (15-24) sont plus mobiles que leurs aînés (25+).

Les données du tableau 2 font ressortir clairement la tendance décroissante des valeurs de ρ . Cette tendance est très bien décrite par un modèle de régression non linéaire, $\rho_i = a + bt + ct^{-1}$. Les valeurs de R^2 (corrélation multiple) se rapprochent de 1 (> 0.98). Les valeurs prédites de ρ_j semblent donc très satisfaisantes. Par ailleurs, Lee (1989a et 1989b) détermine ρ_j en extrapolant $\hat{\rho}_3$ et $\hat{\rho}_4$. La différence entre les valeurs prédites et les valeurs extrapolées est toutfois très mince, étant inférieure à 0.01 pour toutes les caractéristiques sauf UNEMP, UNEMP 15-24 et UNEMP 25+, où elle n'excède pas 0.03.

4.2 Valeurs estimées des coefficients de corrélation γ

Comme nous l'avons mentionné dans la sous-section 3.2, il y a deux façons d'estimer γ_2 , γ_3 et γ_4 , soit par l'équation (3) ou l'équation (4). Nous désignerons la première comme la Méthode 1 et la seconde comme la Méthode 2. Seule la Méthode 1 peut servir à estimer γ_1 , tandis que γ_5 ne peut être estimé directement que par la Méthode 2. Dans le tableau 3, nous comparons les deux méthodes à l'aide de données empiriques. Les valeurs de γ_5 indiquées pour la Méthode 1 ont été prédites au moyen d'un modèle linéaire logarithmique. Le tableau montre que les deux méthodes ont donné des résultats assez différents. On voit très bien que les coefficients calculés à l'aide de la Méthode 2 suivent une tendance à la hausse qui est contraire à notre intuition, tandis que les coefficients calculés par la Méthode 1 sont plus acceptables. En outre, si nous comparons les valeurs estimées du tableau 3 à γ_1 , qui ne peut être calculée que par la Méthode 1, cette dernière semble produire des résultats plus acceptables que la Méthode 2. Nous avons donc opté pour la Méthode 1. Néanmoins, dans des circonstances normales, les deux méthodes devraient être équivalentes et produire des résultats comparables. Il semble, dans ce cas-ci, que les données réelles ne soient pas parfaitement conformes aux hypothèses que nous avons posées pour définir les équations.

Tableau 3
Comparaison de valeurs estimées de γ_2 , γ_3 , γ_4 et γ_5 calculées à l'aide de méthodes différentes (Ontario, 1980-1981)

Caractéristique	Méthode	γ_2	γ_3	γ_4	γ_5
IN LF	1	0.141	0.128	0.133	0.135
	2	0.107	0.105	0.116	0.120
EMP	1	0.136	0.142	0.142	0.147
	2	0.100	0.115	0.126	0.133
EMP AG	1	0.483	0.474	0.486	0.451
	2	0.321	0.370	0.407	0.448
EMP NON-AG	1	0.150	0.147	0.157	0.163
	2	0.117	0.134	0.145	0.149
UNEMP	1	0.074	0.076	0.063	0.080
	2	0.043	0.056	0.046	0.043

Nota: Les méthodes 1 et 2 correspondent aux équations (3) et (4) respectivement (section 3).

4. RÉSULTATS ET ANALYSE

En nous servant des méthodes décrites dans la section précédente, nous avons établi les valeurs estimées de ρ et γ à partir des données de l'EPA de 1980-1981 et de 1985-1987 pour 5 caractéristiques: actif (INLF), occupé (EMP), occupé en agriculture (EMP AG), occupé dans d'autres secteurs (EMP NON-AG), en chômage (UNEMP). Avec les données de 1980-1981, nous avons estimé les corrélations de panel pour trois provinces seulement: la Nouvelle-Ecosse (N.-É.), l'Ontario (ONT) et la Colombie-Britannique (C.-B.). Toutefois, avec des données plus récentes (mars 1985 - février 1987), nous avons établi des estimations pour toutes les provinces. Nous avons aussi ajouté quatre autres caractéristiques, soit les personnes occupées et les personnes en chômage pour deux groupes d'âge, 15-24 et 25+ (EMP 15-24, EMP 25+, UNEMP 15-24, UNEMP 25+). En ce qui concerne l'estimation de τ , nous nous sommes servis uniquement des données de 1985-1987 pour les quatre dernières caractéristiques et pour les provinces de N.-É., d'ONT et d'Alberta (ALT).

Nous ne présentons et ne commentons ici qu'une partie des résultats. Le lecteur trouvera tous les résultats dans Lee (1989b).

4.1 Valeurs estimées des coefficients de corrélation ρ

Le tableau 2 donne les valeurs estimées des coefficients de corrélation ρ . Bien que les données de 1985-1987 aient permis de calculer des valeurs estimées de ρ pour les cinq caractéristiques initiales (INLF, EMP, EMP AG, EMP NON-AG, UNEMP) et pour toutes les provinces, les deux séries d'estimations qui figurent dans le tableau 2 à des fins de comparaison ne portent que sur trois provinces (N.-É., ONT et C.-B.). Ce tableau contient aussi les estimations de ρ pour les quatre autres caractéristiques (EMP 15-24, EMP 25+, UNEMP 15-24, UNEMP 25+) pour la N.-É. et l'Ontario.

Tableau 2

Valeurs estimées du coefficient de corrélation ρ (données de 1980-1981 et de 1985-1987)											
Prov. Caractéristique						Données de 1980-1981					
						Données de 1985-1987					
						ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_5
N.-É.	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.862	0.797	0.744	0.679	0.622	0.845
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.866	0.783	0.714	0.651	0.590	0.863
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.913	0.837	0.756	0.678	0.598	0.912
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.865	0.774	0.710	0.649	0.594	0.873
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.590	0.455	0.333	0.243	0.145	0.703
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.773	0.632	0.556	0.495	0.446	0.779
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.754	0.600	0.528	0.467	0.415	0.754
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.729	0.574	0.502	0.441	0.389	0.729
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.705	0.550	0.478	0.417	0.365	0.705
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.446	0.301	0.229	0.168	0.116	0.446
ONT	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.843	0.782	0.717	0.674	0.622	0.846
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.852	0.779	0.709	0.664	0.611	0.853
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.955	0.926	0.901	0.861	0.827	0.962
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.861	0.791	0.724	0.678	0.625	0.866
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.580	0.445	0.334	0.286	0.222	0.579
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.747	0.605	0.500	0.429	0.356	0.747
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.732	0.588	0.484	0.413	0.340	0.732
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.701	0.550	0.478	0.417	0.365	0.701
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.660	0.500	0.428	0.367	0.315	0.660
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.448	0.303	0.231	0.170	0.118	0.448
C.-B.	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.849	0.767	0.705	0.665	0.622	0.817
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.835	0.755	0.695	0.651	0.607	0.851
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.896	0.809	0.733	0.656	0.582	0.938
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.855	0.769	0.715	0.661	0.616	0.857
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.516	0.407	0.334	0.320	0.294	0.634
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.767	0.625	0.553	0.492	0.440	0.767
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.732	0.588	0.484	0.413	0.340	0.732
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.701	0.550	0.478	0.417	0.365	0.701
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.660	0.500	0.428	0.367	0.315	0.660
	IN LF	EMP	EMP AG	EMP NON-AG	UNEMP	0.448	0.303	0.231	0.170	0.118	0.448

On peut alors calculer la valeur estimée de γ_j à l'aide de l'équation suivante:

(3)
$$\gamma_j = \frac{1}{f} \left[\frac{\text{Cov}(A,B)}{\sqrt{V(A)V(B)}} - (6 - j)\rho_j \right],$$

en ayant recours à des valeurs estimées des éléments du membre de droite. On peut aussi estimer directement les coefficients de corrélation γ , y compris γ_5 , en se servant de l'équation

(4)
$$\gamma_j = \frac{\text{Cov}(A_j,B_j)}{\sqrt{V(A_j)V(B_j)}},$$

où $A_j = \sum_{l=1}^j y_{m,l}$ et $B_j = \sum_{l=7-j}^6 y_{m-j,l}$, $j = 2, \dots, 5$. Dans la section 4, nous comparons les deux méthodes au moyen de données empiriques.

L'équation (4) permet de calculer d'autres coefficients $\gamma(\gamma_j, j = 6, \dots, 10)$; dans ce cas,

$$A_j = \sum_{l=j-5}^6 y_{m,l},$$

$$B_j = \sum_{l=1}^{12-j} y_{m-j,l}.$$

Il n'existe pas de méthode simple pour estimer directement ou indirectement γ_{11} . Les valeurs estimées $\hat{\gamma}_5$ et $\hat{\gamma}_{11}$ ont toutes deux été prédites au moyen d'un modèle linéaire logarithmique, $\gamma = \exp(a + br)$, $r = 1, \dots, 4, 6, \dots, 10$.

3.3 Estimation des coefficients de corrélation τ

On peut estimer τ de la même manière que ρ en remplaçant simplement $y_{m,l}$ par $x_{m,l}$. Soit $A = \sum_{l=j+1}^6 x_{m,l}$ et $B = \sum_{l=7-j}^6 y_{m-j,l}$, $j = 0, 1, \dots, 4$. Nous avons alors

$$\text{Cov}(A,B) = (6 - j) \tau_j \sigma_x \sigma_y,$$

$$V(A) = (6 - j) \sigma_x^2,$$

$$V(B) = (6 - j) \sigma_y^2,$$

ce qui donne

(5)
$$\tau_j = \frac{\text{Cov}(A,B)}{\sqrt{V(A)V(B)}}, \quad j = 0, 1, \dots, 4.$$

L'équation ci-dessus permet d'estimer tous les τ sauf τ_5 , qui est prédit au moyen d'un modèle linéaire logarithmique, $\tau = \exp(a + br)$, $r = 1, \dots, 4$.

3.1 Estimation des coefficients de corrélation ρ

Soit $A = \sum_{l=1}^6 y_{m,l}$ et $B = \sum_{l=1}^6 y_{m-1,l}$. On obtient les valeurs A et B en faisant abstraction, dans le premier cas, du Panel 1 et, dans le second cas, du Panel 6. Notons que les panels éliminés ne sont pas des panels communs alors que tous les autres le sont. À l'aide du programme mentionné plus haut, nous calculons les valeurs estimées de $V(A - B)$, $V(A)$ et de $V(B)$, puis nous nous servons de l'équation (1) pour estimer $\text{Cov}(A, B)$. D'après les hypothèses posées dans la section 2, il est facile de constater que

$$\begin{aligned}\text{Cov}(A, B) &= 5\rho_1\sigma_y^2, \\ V(A) &= V(B) = 5\sigma_y^2,\end{aligned}$$

et, par conséquent,

$$\rho_1 = \frac{\text{Cov}(A, B)}{\sqrt{V(A)V(B)}}. \quad (2)$$

Il ne reste plus qu'à utiliser des valeurs estimées de $\text{Cov}(A, B)$, $V(A)$ et $V(B)$ pour obtenir la valeur estimée de ρ_1 . On peut estimer ρ_2 , ρ_3 et ρ_4 de la même façon en posant $A = \sum_{l=j+1}^6 y_{m,l}$ et $B = \sum_{l=1}^6 y_{m-j,l}$, $j = 2, 3, 4$. Cependant, lorsqu'il s'agit d'estimer ρ_5 de cette manière, des difficultés surgissent. En effet, lorsqu'on élimine tous les panels non communs pour les mois m et $m - 5$, il ne reste plus qu'un panel pour chaque mois et cela complique l'estimation de la variance pour ce qui a trait aux unités autoreprésentatives (UAR). Les UAR représentent les grandes agglomérations et chacune de ces unités fait l'objet d'un échantillonnage indépendant. Le problème ne se pose pas pour les unités non autoreprésentatives (UNAR), qui représentent les régions situées à l'extérieur des UAR, notamment les régions rurales et les petits centres urbains. Chaque unité primaire d'échantillonnage (UPÉ) qui sert d'échantillon répété pour l'estimation de la variance compte tous les groupes de renouvellement; par conséquent, même si on élimine 5 panels non communs, il y a toujours un panel de représenté dans l'UPÉ, de sorte qu'il est possible de calculer la variance. Dans les UAR toutefois, les groupes de renouvellement tiennent lieu d'échantillons répétés et s'il ne reste plus qu'un groupe, il n'y a donc qu'un échantillon répété par strate et on ne peut, par conséquent, calculer la variance de la manière habituelle. Nous avons donc calculé ρ_5 par prédiction en utilisant le modèle de régression non linéaire $\rho = a + bt + ce^{-t}$, $t = 1, \dots, 4$. Dans la sous-section 4.1, nous verrons une autre façon d'estimer ρ_5 .

3.2 Estimation des coefficients de corrélation γ

On constate facilement que $\text{Cov}(A, B) = (5\rho_1 + \gamma_1)\sigma_y^2$ si $A = \sum_{l=1}^6 y_{m,l}$ et $B = \sum_{l=1}^6 y_{m-1,l}$. En règle générale,

$$\text{Cov}(A, B) = \{(6 - j)\rho_j + j\gamma_j\}\sigma_y^2,$$

où

$$A = \sum_{l=1}^6 y_{m,l},$$

$$B = \sum_{l=1}^6 y_{m-j,l}, \quad j = 1, \dots, 4.$$

Voici donc les définitions formelles de ρ_j de γ_j et de τ_j :

Soit $y_{m,l}$ la valeur estimée d'une caractéristique d'intérêt, établie à partir du panel $P(m,l)$ de l'EPA. Nous supposons que $V(y_{m,l}) = \sigma_y^2$, quels que soient m et l . Alors, ρ_j est défini au moyen de l'équation

$$\text{Cov}(y_{m,l}, y_{m-j,l-j}) = \rho_j \sigma_y^2, \quad 1 \leq j \leq 5, \quad j < l \leq 6,$$

et γ_j , au moyen de l'équation

$$\text{Cov}(y_{m,l}, y_{m-j,l+j}) = \gamma_j \sigma_y^2,$$

où $1 \leq l \leq j$ si $1 \leq j \leq 6$ et $j - 5 \leq l \leq 6$ si $7 \leq j \leq 11$.

Il serait normal de supposer que ρ_j et γ_j diminuent lorsque j augmente et que ρ_j est plus grand que γ_j étant donné que le premier désigne une corrélation qui met en rapport des ménages voisins tandis que le second désigne une corrélation qui met en rapport des ménages entre un panel et le panel qui précède son panel antécédent (appelé *second panel antécédent* et identifié par des parenthèses doubles dans le tableau 1) en désignant cette corrélation par δ_j ; nous avons ainsi $\delta_7, \delta_8, \dots, \delta_{17}$. Les valeurs de δ seront moins élevées que celles de γ_j mais pourraient s'en rapprocher sensiblement, pour le même indice, à cause de la proximité géographique du panel antécédent et du second panel antécédent. Toutefois, compte tenu du temps et des ressources dont nous disposons, nous n'étudierons pas ici les coefficients de corrélation δ . Nous supposons que $\text{Cov}(y_{m,l}, y_{m,l'}) = 0$ si $l \neq l'$ et que $\text{Cov}(y_{m,l}, y_{m-j,l'}) = 0$ si $P(m-j, l')$ et $P(m, l)$ ne sont pas des panels communs et si le premier n'est pas non plus le panel antécédent du second.

Afin de définir le coefficient de corrélation τ_j , posons $x_{m,l}$ comme la valeur estimée d'une autre caractéristique, établie à partir du panel $P(m, l)$ de l'EPA, et posons comme hypothèse que $V(x_{m,l}) = \sigma_x^2$ est indépendante de m et de l . Alors, le coefficient de corrélation τ_j peut être déterminé à l'aide de l'équation

$$\text{Cov}(y_{m,l}, y_{m-j,l-j}) = \tau_j \sigma_x \sigma_y, \quad 0 \leq j \leq 5, \quad j < l \leq 6.$$

3. ESTIMATION DES COEFFICIENTS DE CORRÉLATION DE PANEL

Comme nous disposons d'un programme d'ordinateur pour l'estimation de la variance, la méthode décrite ci-dessous a été élaborée de manière que nous puissions utiliser ce programme sans devoir y apporter trop de modifications. La méthode utilisée en l'occurrence est la méthode généralisée de Keyfitz (Choudhry et Lee 1987; Lee 1989a), mieux connue sous le nom de méthode de Taylor. Le programme peut calculer des estimations de la variance de combinaisons linéaires de valeurs estimées mensuelles.

Pour estimer les coefficients de corrélation voulus à l'aide du programme mentionné plus haut, nous utilisons l'équation fondamentale suivante:

$$\text{Cov}(A, B) = \frac{V(A) + V(B) - V(A - B)}{2}. \quad (1)$$

$V(A - B)$, $V(A)$ et $V(B)$ peuvent être déterminées à l'aide du programme et $\text{Cov}(A, B)$ peut être calculée au moyen de l'expression ci-dessus. On trouvera dans Kish (1965) une expression pour $V(A - B)$ dont on peut déduire l'équation (1).

2. DÉFINITION DES COEFFICIENTS DE CORRÉLATION DE PANEL

Pour définir divers types de corrélation de panel, il faut tout d'abord définir la notion de "panels communs" et de "panel antécédent". Un panel est identifié par un numéro qui indique depuis combien de mois ce panel est présent dans l'échantillon. Ainsi, Panel 1 au mois m devient Panel 2 au mois $m + 1$, Panel 3 au mois $m + 2$, et ainsi de suite. On utilise souvent le terme groupe de renouvellement pour désigner un panel, peu importe depuis combien de mois ce panel est présent dans l'échantillon. Par exemple, le groupe de renouvellement 1 introduit dans l'échantillon en janvier est toujours identifié comme le groupe 1 jusqu'à ce qu'il soit supprimé de l'échantillon, en juillet. Ainsi, Panel 1 en janvier désigne le groupe de renouvellement 1 et Panel 2 en février désigne le même groupe de renouvellement, qui en est à son deuxième mois dans l'échantillon, et ainsi de suite.

Deux panels qui représentent le même groupe de renouvellement à des mois différents sont désignés comme des *panels communs*. Lorsqu'un groupe de renouvellement est supprimé de l'échantillon, il est remplacé le plus souvent par un groupe formé de ménages voisins et auquel on attribue le même numéro de renouvellement. Le panel associé au groupe de renouvellement que l'on supprime est le *panel antécédent* à celui qui est associé au nouveau groupe. Ainsi, dans l'exemple ci-dessus, Panel 6 en juin, qui représente le groupe de renouvellement 1, est le panel antécédent au Panel 1 de juillet. Le tableau 1 montre schématiquement les panels communs et les panels antécédents pour les mois donnés m et $m - j$.

Puisqu'on peut identifier chaque panel par deux éléments (mois et numéro de panel), désignons par $P(m)$ (mois, numéro de panel) un panel quelconque. Alors, $P(m, 4)$ et $P(m - 1, 3)$, par exemple, sont des panels communs à 1 mois d'intervalle. De même, $P(m, 4)$ et $P(m - 2, 2)$, sont des panels communs à 2 mois d'intervalle. On désigne par ρ_j le coefficient de corrélation des valeurs estimées d'une caractéristique établies à partir de panels communs à j mois d'intervalle. Évidemment, il ne peut exister de panels communs à plus de 5 mois d'intervalle; la valeur de l'indice j ne peut donc excéder 5. Nous supposons que ρ_j est indépendant de m et du numéro de panel. En revanche, il est fonction de j et varie selon les caractéristiques.

On désigne par γ_j le coefficient de corrélation des valeurs estimées d'une caractéristique établies à partir d'un panel donné et de son panel antécédent, séparés par j mois d'intervalle. Dans ce cas toutefois, la valeur de j peut atteindre 11; autrement dit, γ_{11} est le dernier coefficient de corrélation de la série et il désigne la corrélation entre les estimations de $P(m, 6)$ et celles de $P(m - 11, 1)$. Nous supposons là aussi que γ est indépendant de m et du numéro de panel. En revanche, comme le coefficient précédent, il est fonction de j et varie selon les caractéristiques. Le troisième type de corrélation de panel est la corrélation entre les valeurs estimées de deux caractéristiques différentes, établies à partir de panels communs à j mois d'intervalle; ce type de corrélation est désignée par τ_j . Dans ce cas, la valeur de l'indice j peut varier de 0 à 5 et nous posons les mêmes hypothèses que pour ρ et γ .

Tableau 1
Panels communs et panels antécédents pour les mois m et $m - j$

m	$m - 1$	$m - 2$	$m - 3$	$m - 4$	$m - 5$	$m - 6$	$m - 7$	$m - 8$	$m - 9$	$m - 10$	$m - 11$
1	(6)	(5)	(4)	(3)	(2)	(1)	(6)	(5)	(4)	(3)	(2)
2	1	(6)	(5)	(4)	(3)	(2)	(1)	(6)	(5)	(4)	(3)
3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	(6)	(5)	(4)
4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	(6)	(5)
5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	(6)
6	5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)

Nota: Les parenthèses simples désignent des panels antécédents et les doubles, des seconds panels antécédents.

Estimation des coefficients de corrélation de panel pour l'Enquête sur la population active du Canada

HYUNSHIK LEE¹

RÉSUMÉ

L'enquête sur la population active du Canada repose sur un plan avec renouvellement de panel. À chaque mois, on renouvelle un sixième de l'échantillon global et on en conserve les cinq sixièmes. Ainsi, une fois qu'un panel est introduit dans l'échantillon, il y demeure 6 mois avant d'en être exclu. Cette caractéristique du plan de sondage ainsi que le mode de sélection des panels font que les estimations de panel pour le même mois ou des mois différents sont corrélées. La corrélation entre deux estimations de panel est désignée comme la corrélation de panel. Nous définissons trois types de corrélation de panel dans cet article: (1) la corrélation (désignée par ρ) entre des estimations de la même caractéristique tirées du même panel à des mois différents; (2) la corrélation (désignée par γ) entre des estimations de la même caractéristique tirées de panels géographiquement rapprochés l'un de l'autre à des mois différents; (3) la corrélation (désignée par τ) entre des estimations de caractéristiques différentes tirées du même panel dans le même mois ou dans des mois différents. En deuxième lieu, nous décrivons des méthodes permettant d'estimer les corrélations de panel et calculons les coefficients de corrélation estimés pour certaines variables en nous servant de données pour 1980-1981 et 1985-1987; nous terminons par une analyse des résultats.

MOTS CLÉS: Enquête par panel; renouvellement; méthode de Taylor.

1. INTRODUCTION

L'enquête sur la population active (EPA) est une enquête permanente menée auprès des ménages à chaque mois et qui repose sur un plan avec renouvellement de panel. L'échantillon de l'EPA est constitué de six panels de même taille qui, à tour de rôle, font place à un nouveau panel à chaque mois. Le nouveau panel passe six mois dans l'échantillon avant d'en être supprimé. (Pour une description détaillée de la méthodologie de l'EPA, voir Platek et Singh (1976) et Singh et coll. (1990).) Par conséquent, il existe une forte corrélation entre les estimations tirées du même panel, donc des mêmes unités d'échantillonnage, à des mois différents. En outre, les panels qui sont supprimés de l'échantillon sont remplacés habituellement par un panel voisin. La proximité géographique des panels fait aussi que les estimations respectives sont corrélées. On appelle ce genre de corrélations des corrélations de panel. Dans cet article, nous voyons comment on peut estimer ces corrélations et présentons les coefficients de corrélation estimés pour certaines variables. Cette étude avait pour but initialement d'analyser la méthode d'estimation composite; ses résultats peuvent néanmoins s'appliquer à toute situation où intervient la corrélation de panel.

Notre article se divise comme suit. Dans la section 2, nous présentons les définitions, les symboles et les hypothèses nécessaires. Dans la section suivante, nous décrivons les méthodes d'estimation pertinentes. Enfin, la section 4 renferme les résultats de l'estimation ainsi qu'une analyse.

¹ H. Lee, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11^{ème} étage, Immeuble R. H. Coats, Parc Tunney, Ottawa, Ontario, K1A 0T6.

- BANDYOPADHYAY, S., CHATTOPADHYAY, A.K., et KUNDU, S.C. (1977). On estimation of population total. *Sankhyā*, Sér. C, 39, 28-42.
- BANDYOPADHYAY, S. (1980). Improved ratio and product estimators. *Sankhyā*, Sér. C, 42, 45-49.
- CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-396.
- GODAMBE, V.P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, Sér. B, 17, 269-278.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- YATES, F., et GRUNDY, P.M. (1953). Selections without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Sér. B, 15, 253-261.

BIBLIOGRAPHIE

3. ÉTUDE EMPIRIQUE DU BIAIS ET DE L'ERREUR QUADRATIQUE MOYENNE

X	0.1	0.2	0.3	0.4
Y	0.5	1.2	2.1	3.2
Population A				
	0.1	0.2	0.3	0.4
Population B				
	0.8	1.4	1.8	2.0
Population C				
	0.1	0.2	0.3	0.4
	0.2	0.6	0.9	0.8

Populations:	A	B	C
Biais relatif de $R(s)$	0.02456	- 0.02785	- 0.00496
Biais relatif de $R^{HT}(s)$	- 0.00379	0.00552	0.00232
EQM de $R(s)$	0.2946	0.2946	0.0824
EQM de $R^{HT}(s)$	0.3159	0.3642	0.0690
Efficacité de $R(s)$ par rapport à $R^{HT}(s)$	1.0723	1.2362	0.8374

REMERCIEMENTS

L'auteur remercie sincèrement les arbitres qui, par leurs commentaires judicieux et constructifs, ont contribué à la version définitive de cet article.

Théorème principal. Étant donné un plan d'échantillonnage symétrique, le rapport de deux totaux non pondérés est un estimateur du rapport de population correspondant dans la mesure où il est défini comme le rapport d'un estimateur sans biais à un autre estimateur sans biais, sauf que le rapport estimé fait abstraction des probabilités de sélection des unités de la population.

Il convient de préciser que dans le cas de plans symétriques, il n'est pas nécessaire que les probabilités de sélection des unités soient égales. Les plans d'échantillonnage symétriques ne sont donc pas nécessairement autopondérés. Les plans autopondérés exigent que $\alpha_i/p(s)$ ait la même valeur pour tous i et s , alors que cette condition n'est pas caractéristique des plans symétriques.

Dans le cas de plans non symétriques, l'équation (2.2) est facile à résoudre puisque les α_i sont faciles à calculer dans la majorité des cas et qu'il n'est pas nécessaire de calculer les probabilités de sélection.

En ce qui concerne l'échantillonnage de n unités sans remise, on compte $\binom{n-1}{N-1}$ échantillons (non ordonnés) qui renferment une unité U_i , donné, de sorte que $\alpha_i = \binom{n-1}{N-1}$ pour tous i et par conséquent, le plan d'échantillonnage avec PPTSR est symétrique. Il convient de souligner qu'il n'y a pas toujours $\binom{n}{N}$ échantillons possibles avec les plans PPTSR. Comme l'indique Connor (1966), il arrivera que dans des plans d'échantillonnage systématique avec PPT (randomisé ou non), un ensemble de n unités ait une probabilité de sélection nulle. Les résultats de notre analyse s'appliquent lorsque le plan PPTSR est tel que, pour aucun ensemble de n unités, il n'existe de probabilité de sélection composée nulle.

Pour ce qui a trait à l'échantillonnage de n unités avec remise, on compte N^n échantillons (ordonnés), de sorte que $\alpha_i = nN^{n-1}$ pour tous i et par conséquent, le plan PPTAR est symétrique.

Dans le cas d'un échantillonnage par strate dans k strates, avec PPTSR, la valeur α pour chaque unité de la strate j est

$$\alpha_j = \frac{n_j}{N_j} \prod_{i=1}^k \left(\frac{n_i}{N_i} \right)$$

qui devient une constante lorsque la répartition est proportionnelle et qu'il n'existe de probabilité de sélection composée nulle pour aucun ensemble d'unités dans aucune strate, N_j et n_j étant respectivement l'effectif de la population et de l'échantillon pour la strate $j = 1, 2, \dots, k$. On peut aussi avoir des plans à plusieurs degrés symétriques en appliquant le même mode de répartition.

Dans le cas de l'échantillonnage avec PPTAR, soulignons que l'estimateur non biaisé de $T(Y)$ défini en (2.1) n'est pas acceptable. On peut améliorer cet estimateur en remplaçant $n(i,s)$ et α_i par $n^*(i,s)$ et α_i^* , respectivement, où $n^*(i,s)$ est égal à 1 si $n(i,s)$ est au moins égal à 1, et égal à 0 si $n^*(i,s)$ est nul, et où α_i^* est défini par rapport à $n^*(i,s)$. En l'occurrence, $\alpha_i^* = N^n - (N^n - 1)^n$, soit le nombre d'échantillons (ordonnés) contenant une unité U_i donnée. En ce qui a trait à la comparaison d'estimateurs, il n'est pas possible d'obtenir une expression en forme analytique fermée pour l'efficacité relative, même lorsqu'il s'agit de plans PPTAR.

Pour pouvoir comparer les biais relatifs et l'efficacité d'estimateurs, nous proposons une étude empirique réalisée au moyen d'un plan PPTSR. Une autre formule tout aussi intéressante serait d'analyser la variance et le biais de gros échantillons au moyen du développement de Taylor.

Il est clair que nous ne pouvons estimer la variance de $R(s)$ sans connaître les poids d'échantillonnage ou sans poser d'autres hypothèses. Cependant, si s_1 et s_2 sont deux demi-échantillons prélevés au moyen du même plan symétrique (comme deux échantillons PPTSR indépendants de même taille), R est estimé au moyen de la formule $[R(s_1) + R(s_2)]/2$, et l'estimateur non biaisé de la variance de l'estimateur est $[R(s_1) - R(s_2)]^2/4$.

2. PLANS D'ÉCHANTILLONNAGE SYMÉTRIQUES

Considérons une population finie composée de N unités U_1, U_2, \dots, U_N . Désignons par Y_i et X_i , les valeurs de deux variables d'intérêt, Y et X qui se rapportent à l'unité $U_i, i = 1, 2, \dots, N$. Le problème consiste à estimer un taux ou ratio $R = T(Y)/T(X)$ où $T(Y) = Y_1 + Y_2 + \dots + Y_N$, et $T(X)$ est défini de la même façon avec la variable X . La procédure normale est de calculer des estimations sans biais de $T(Y)$ et $T(X)$ puis d'utiliser le rapport de ces estimations comme estimateur de R . Dans cet article, nous nous proposons d'appliquer la même méthode de telle manière que le rapport des estimations fasse abstraction des probabilités de sélection des unités d'échantillonnage.

Définissons un plan d'échantillonnage. Soit S l'ensemble de tous les échantillons possibles tel que $p(s) > 0$, où $p(s)$ est la probabilité de tirer l'échantillon s , et $\sum_{s \in S} p(s) = 1$. Pour s dans S et $i = 1, 2, \dots, N$,

$n(i, s)$ = le nombre de fois que U_i est incluse dans s , et $\alpha_i = \sum_{s \in S} n(i, s)$, le nombre de fois que U_i est incluse dans tous les échantillons possibles.

$S, p(s)$, et α_i dépendent du plan d'échantillonnage.

Définition 2.1. Un plan d'échantillonnage est dit symétrique si $\alpha_i = \alpha$, pour tous $i = 1, 2, \dots, N$.

L'estimateur ci-dessous, qui est fondé sur l'échantillon s , et qui appartient à la catégorie d'estimateurs linéaires sans biais de Godambe (1955) pour $T(X)$, a été analysé par Bandyopadhyay et coll. (1977).

$$(2.1) \qquad T(X, s) = \sum_N^{i=1} Y_i n(i, s) \alpha_i^{-1} p^{-1}(s).$$

De toute évidence, $T(Y, s)$ est un estimateur non biaisé de $T(Y)$. Un estimateur du ratio $R = T(Y)/T(X)$, fondé sur un échantillon s et défini comme le rapport d'un estimateur sans biais de $T(Y)$ à un estimateur sans biais de $T(X)$, est

$$(2.2) \qquad R(s) = T(Y, s)/T(X, s) = \sum_N^{i=1} Y_i n(i, s) \alpha_i^{-1} \bigg/ \sum_N^{i=1} X_i n(i, s) \alpha_i^{-1}.$$

Dans le cas de plans d'échantillonnage symétriques, $\alpha_i = \alpha$ pour tous i et (2.2) devient

$$(2.3) \qquad R(s) = \sum_N^{i=1} Y_i n(i, s) \bigg/ \sum_n X_i n(i, s) = \frac{\text{total non pondéré des valeurs } Y \text{ dans l'échantillon}}{\text{total non pondéré des valeurs } X \text{ dans l'échantillon}}.$$

Les observations précédentes sont résumées dans le théorème suivant.

Estimation d'un rapport sans connaître le plan d'échantillonnage

SHIBDAS BANDYOPADHYAY¹

RÉSUMÉ

L'auteur montre que, pour une catégorie d'estimateurs linéaires sans biais appliqués à une catégorie de plans d'échantillonnage, il est possible d'estimer un rapport de population au moyen d'un rapport d'estimateurs non biaisés sans connaître les poids d'échantillonnage. Cette catégorie de plans d'échantillonnage comprend des plans aussi courants que l'échantillonnage avec probabilités inégales avec ou sans remise, l'échantillonnage stratifié avec répartition proportionnelle et probabilités inégales de sélection sans remise dans chaque strate, etc.

MOTS CLÉS: Rapport de totaux non pondérés; échantillonnage symétrique.

1. INTRODUCTION

Soit m le nombre d'adultes qui savent lire et écrire parmi t adultes dans un échantillon de n familles tiré d'une population donnée. Nous supposons que le taux d'alphabétisation des adultes dans la population R est estimé au moyen de la formule $r = m/t$. De même, pour un tableau à double entrée qui donne la répartition (en pourcentage) d'individus selon l'âge et le sexe, nous supposons qu'une fréquence de case est estimée au moyen du rapport entre le nombre d'individus dans la case et le nombre total d'individus pour l'échantillon de n familles (ce rapport étant multiplié par 100 pour obtenir un pourcentage).

Peu importe la méthode de sélection des familles, un rapport formé de deux totaux non pondérés et qui vise à estimer un ratio ou une répartition en pourcentage répond aux besoins de nombreux utilisateurs non spécialistes. De fait, certains rapports d'enquêtes renferment des tableaux dont les éléments sont estimés de cette façon, comme si le plan d'échantillonnage était un plan autopondéré.

En revanche, si les n familles devaient être choisies suivant un plan d'échantillonnage (à un degré) avec PPTSR, il faudrait normalement déterminer des totaux pondérés pour obtenir des estimateurs non biaisés des numérateurs et des dénominateurs correspondants avant de calculer un rapport ou une répartition en pourcentage.

Dans cette étude, nous montrons que, pour des plans d'échantillonnage comme le plan à un degré avec PPTSR et sans aucune autre hypothèse,

i) un rapport formé de deux totaux non pondérés peut servir à estimer le rapport de population correspondant; le premier doit être vu comme le rapport d'un *estimateur sans biais* du numérateur à un *estimateur sans biais* du dénominateur;

ii) il existe une catégorie de plans d'échantillonnage, qui ne sont pas autopondérés, pour lesquels la proposition (i) se vérifie. Cette catégorie comprend les plans d'échantillonnage à un degré avec probabilités inégales avec ou sans remise et les plans d'échantillonnage stratifié avec répartition proportionnelle et probabilités inégales de sélection sans remise dans chaque strate.

¹ Shibdas Bandyopadhyay, Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, Inde.

4. CONCLUSION

Dans cet article, nous avons vu comment améliorer des estimations d'enquête infra-annuelles au moyen d'estimations d'enquête annuelles. Nous avons présenté un moyen simple et inédit d'établir une série chronologique. La méthode en question peut être appliquée à l'aide d'un programme d'informatique de sorte que son exécution soit automatisée. Elle se distingue avant-tout des méthodes plus classiques (ex.: méthode de Denton) par le fait qu'elle tient compte des erreurs d'échantillonnage. Par la même occasion, nous avons abordé certaines questions liées à l'utilisation de cette méthode. Deux grandes questions pratiques ont été soulevées: l'établissement d'un tableau de données chronologiques et l'établissement préliminaire. Nous devons chercher des façons d'aborder ces deux questions.

BIBLIOGRAPHIE

- BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day.
- CHOLETTE, P.A. (1988a). Benchmarking and Interpolation of Time Series. Statistics Canada, Working Paper No. TSRA-87-014E.
- CHOLETTE, P.A. (1988b). Benchmarking Systems of Socio-Economic Time Series. Statistics Canada, Document de travail N° TSRA-88-017E.
- CHOLETTE, P.A., et DAGUM, E.B. (1989). Benchmarking Socio-Economic Time Series Data: A Unified Approach. Document de travail N° TSRA-89-006E, Statistique Canada.
- DENTON, F.T. (1971). Adjustment on Monthly or Quarterly Series to Annual Totals: An approach Based on Quadratic Minimization. *Journal of the American Statistical Association*, 66, 99-102.
- HILLMER, S.C., et TRABELSI, A. (1987). Benchmarking of Economic Time Series. *Journal of the American Statistical Association*, 82, 1604-1071.
- LANIEL, N., et FYFE, K. (1989). Benchmarking of Economic Time Series. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.
- LANIEL, N., et FYFE, K. (1990). Benchmarking of Economic Time Series. Méthodes d'enquêtes-entreprises, Statistique Canada, Document de travail du Projet de remaniement des enquêtes-entreprises.
- MONSOUR, N.J., et TRAGER, M.L. (1979). Revision and Benchmarking of Business Time Series. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

Dans la figure 3.1, les estimations mensuelles sont représentées par la ligne continue et les estimations annuelles, par les traits horizontaux. Ces traits correspondent aux quotients des estimations mensuelles étalonnées au moyen de la méthode fondée sur le modèle pour niveaux, tandis que la ligne en pointillés représente la série d'estimations mensuelles étalonnées à l'aide de la méthode de Denton.

Nous voyons d'après cette figure que la série étalonnée au moyen de la méthode fondée sur le modèle pour niveaux a le même taux de changement annuel que la série mensuelle originale, tandis que celle étalonnée au moyen de la méthode de Denton suit plutôt le mouvement des estimations annuelles. Nous constatons aussi que les deux séries étalonnées sont au-dessus de la série originale.

La différence de mouvement d'une année à l'autre entre les deux séries étalonnées peut s'expliquer de la façon suivante. Selon la méthode fondée sur le modèle pour niveaux, on obtient le mouvement de la série étalonnée en pondérant les estimations annuelles et infra-annuelles avec l'inverse des variances d'échantillonnage de ces estimations. Comme, dans cet exemple, les estimations infra-annuelles sont beaucoup plus précises que les estimations annuelles, la série étalonnée suit le mouvement des estimations mensuelles. En revanche, dans le cas de la méthode de Denton, le mouvement de la série étalonnée ne peut être axé que sur celui de la série d'estimations annuelles, peu importe le degré de précision de ces estimations. En ce sens, la méthode fondée sur le modèle pour niveaux est supérieure à la méthode de Denton.

Enfin, le fait que les deux séries étalonnées se trouvent au-dessus de la série originale confirme simplement que les deux méthodes renferment un mécanisme de correction pour le biais des estimations mensuelles.

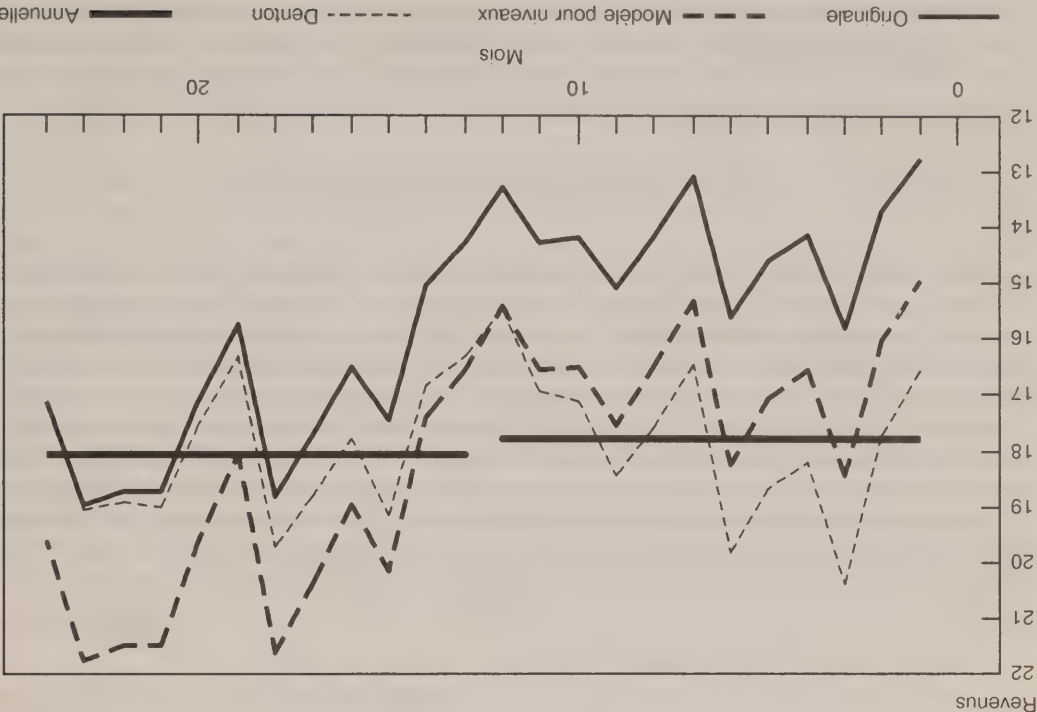


Figure 3.1 La série originale, les deux séries étalonnées, et la série annuelle

(3.6)
$$y_t = \alpha \theta_t + u_t \quad t = 1, 2, \dots, n,$$

où α est un paramètre fixe qui tient compte du biais relatif constant et u_t est défini comme dans l'équation (3.3). Les estimations annuelles sont expliquées par le modèle (3.4). On obtient les estimations étalonnées en appliquant les moindres carrés aux modèles ci-dessus. L'algorithme pertinent est le même que celui utilisé pour la méthode décrite en 3.3.

3.5 Analyse

De toutes les méthodes exposées ci-dessus, la dernière est celle qui se prête le mieux à l'étalonnage d'une série chronologique tirée d'une grande enquête. Elle offre une base statistique qui permet de calculer des régions de confiance et de tester la validité de l'ajustement du modèle étalonné. Le choix du test doit être exercé avec soin puisque les estimations étalonnées, θ_t , ont un très petit nombre de degrés de liberté, $m - 1$ (nombre d'observations annuelles moins un), par rapport au nombre d'observations, $n + m$. De plus, le faible nombre de degrés de liberté donne à penser que le modèle pour niveaux vraisemblablement des estimations étalonnées qui afficheront le même profil chronologique que les estimations infra-annuelles. Une question pratique que soulèvent actuellement les méthodes d'étalonnage qui tiennent compte des erreurs d'échantillonnage, comme celle décrite dans la section 3.4, est le calcul des covariances d'échantillonnage de deux estimations de niveau se rapportant à deux périodes différentes. Devrait-on modéliser ces covariances ou les calculer directement à l'aide du plan de sondage pour toutes les paires de périodes? Du point de vue théorique, il est préférable de les calculer directement puisque la suite des erreurs d'échantillonnage est en soi un processus stochastique non stationnaire étant donné que les variances-covariances de la population varient dans le temps. Toutefois, comme le calcul de l'ensemble des covariances d'échantillonnage peut s'avérer une entreprise laborieuse, il faut continuer de chercher une solution au problème du calcul des covariances d'échantillonnage.

3.6 Exemple

Afin de comparer la méthode de Denton, décrite dans la section 3.1, à celle qui repose sur le modèle pour niveaux, proposée dans la section 3.4, nous avons utilisé l'une et l'autre pour résoudre un problème d'étalonnage particulier et intéressant, où la variance d'échantillonnage des estimations annuelles est six fois plus élevée que la variance d'échantillonnage des estimations mensuelles correspondantes. Cet exemple fait ressortir clairement, pour ce cas-ci, la supériorité de la méthode fondée sur le modèle pour niveaux par rapport à la méthode de Denton. Bien que ce problème soit observable en pratique, nous avons eu recours ici à des données simulées. Premièrement, nous avons tiré vingt-quatre estimations mensuelles d'une enquête économique existante, puis avons défini arbitrairement une matrice de covariances d'échantillonnage pour ces estimations. Les variances et les covariances ont été calculées au moyen d'un coefficient de variation unique pour toutes les périodes et du profil de corrélation suivant:

$$\rho_{ij} = 1 - \frac{24}{|j - i|} \quad \text{pour } i = 1, 2, \dots, 24 \text{ et } j = 1, 2, \dots, 24$$

où i et j sont les indices d'un couple d'estimations mensuelles. Ensuite, nous avons construit deux estimations annuelles correspondantes selon les critères suivants: dans le premier cas, une estimation 25% plus élevée que la somme des douze premières estimations mensuelles; dans le second cas, une estimation seulement 5% plus élevée que la somme des douze dernières estimations mensuelles. Enfin, pour ces deux estimations annuelles, nous avons défini des variances d'échantillonnage six fois plus élevées que celles des estimations mensuelles correspondantes et avons fixé leur coefficient de corrélation à 0.5.

Comme l'indiquent les modèles ci-dessus, la méthode de Hillmer et Trabelsi tient compte des variances et covariances d'échantillonnage des estimations annuelles et infra-annuelles. Malheureusement, elle ne tient pas compte des erreurs systématiques que peuvent renfermer les estimations infra-annuelles. En outre, comme il s'agit de modèles ARMMI, il serait trop coûteux d'utiliser cette méthode pour de grandes enquêtes d'où sont tirées des centaines de séries de données. Il serait donc plus profitable de n'utiliser ce type de méthode que pour un petit nombre d'indicateurs économiques majeurs. Si les modèles ARMMI sont mal spécifiés, on risque par ailleurs d'obtenir un lissage excessif des données.

Cholette et Dagum (1989) ont modifié la méthode de Hillmer et Trabelsi en utilisant un modèle "d'intervention" au lieu d'un modèle ARMMI. Cela permet de modéliser les effets systématiques contenus dans la série chronologique. Selon les auteurs toutefois, cette méthode présente les mêmes lacunes que la méthode de Hillmer et Trabelsi.

3.3 Modèle pour tendances

La méthode suivante a été élaborée dans le but de satisfaire aux conditions d'étalonnage des enquêtes économiques. Elle repose sur l'hypothèse que les estimations infra-annuelles sont expliquées par le modèle:

$$y_t = \theta_t \frac{y_{t-1}}{\theta_{t-1}} + v_t \quad t = 1, 2, \dots, n \tag{3.5}$$

et les estimations annuelles, par le modèle (3.4), où:

$\{\theta_t\}$ est une suite de paramètres infra-annuels (valeurs réelles), comme dans la méthode de Denton,
 $\{v_t\}$ est une suite d'erreurs d'échantillonnage infra-annuelles corrélées des tendances, avec comme vecteur de moyennes et matrice de covariances $(0, \Sigma_v)$.

On obtient les estimations étalonnées en appliquant les moindres carrés aux modèles ci-dessus. L'algorithme de Gauss-Newton utilisé à cette fin ainsi que le calcul de la matrice des covariances des estimations étalonnées sont décrits dans Laniel et Fyfe (1989) ou (1990).

On peut utiliser cette méthode lorsque les données repères proviennent d'un recensement ou d'une enquête annuelle avec échantillons chevauchants et lorsque les estimations de niveau infra-annuelles sont biaisées, à la condition que le biais relatif soit fixe. En pratique, l'hypothèse d'un biais relatif constant se vérifiera si les opérations de mise à jour de la base de sondage se font à un rythme régulier, c'est-à-dire lorsque la proportion d'unités absentes de la base de sondage se stabilise avec les années. En outre, cette hypothèse suppose que les entreprises non énumérées ont le même comportement que celles qui figurent dans la base de sondage. La méthode décrite ci-dessus pose toutefois un problème technique. En effet, on ne peut calculer directement la matrice des variances-covariances d'échantillonnage des tendances infra-annuelles et il faut donc recourir à une formule d'approximation. Après avoir utilisé l'approximation de Taylor du premier degré, on a constaté que dans certains cas, les variances et covariances d'échantillonnage étaient nulles ou négatives alors qu'elles devraient être positives. C'est pourquoi nous présentons dans la section suivante un modèle qui peut remplacer avantageusement le modèle (3.5).

3.4 Modèle pour niveaux

La méthode ci-dessus est une solution de remplacement pour la méthode précédente et est proposée dans le but d'obtenir plus facilement la matrice des variances-covariances d'échantillonnage des estimations infra-annuelles. Elle suppose que les estimations infra-annuelles sont expliquées par le modèle:

$$(3.1) \qquad \frac{y_t}{\theta_{t-1}} = \frac{y_{t-1}}{\theta_{t-1}} + \epsilon_t, \qquad t = 1, 2, \dots, n$$

sujet à la restriction sur les données annuelles:

$$(3.2) \qquad z_T = \sum_{t \leq T} \theta_t, \qquad T = 1, 2, \dots, m,$$

où:

t désigne une période infra-annuelle,

T désigne une période annuelle,

$\{y_t\}$ est une suite d'estimations biaisées des paramètres infra-annuels (niveaux),

$\{\theta_t\}$ est une suite de paramètres infra-annuels fixes (valeurs réelles des niveaux),

$\{\epsilon_t\}$ est une suite d'erreurs non corrélées et identiquement distribuées avec comme vecteur

de moyennes et matrice de covariances $(\mathbf{0}, \sigma^2 I)$,

$\{z_T\}$ est une suite de données repères annuelles.

Pour obtenir les estimations étalonées, on applique les moindres carrés au modèle défini ci-dessus.

Précisons que la méthode de Denton suppose que le biais suit une marche aléatoire et que les données annuelles, aussi bien qu'infra-annuelles, ne sont entachées d'aucune erreur d'échantillonnage. Malheureusement, il est peu probable que ces hypothèses se vérifient avec des séries économiques (voir section 2).

3.2 Méthode de Hillmer et Trabelsi

En 1987, Hillmer et Trabelsi ont proposé une méthode d'étalement fondée sur les modèles ARMMI de Box et Jenkins (1976). Ils ont supposé que les estimations infra-annuelles étaient expliquées par le modèle:

$$(3.3) \qquad y_t = \theta_t + u_t \qquad t = 1, 2, \dots, n$$

et les estimations annuelles, par le modèle:

$$(3.4) \qquad z_T = \sum_{t \leq T} \theta_t + a_T \qquad T = 1, 2, \dots, m,$$

où:

$\{\theta_t\}$ est une suite de paramètres infra-annuels stochastiques (valeurs réelles des niveaux) qui suivent un modèle ARMMI,

$\{y_t\}$ est une suite d'estimations non biaisées des paramètres infra-annuels,

$\{u_t\}$ est une suite d'erreurs d'échantillonnage infra-annuelles corrélées avec comme vecteur

de moyennes et matrice de covariances $(\mathbf{0}, \Sigma_u)$,

$\{z_T\}$ est une suite d'estimations annuelles non biaisées,

$\{a_T\}$ est une suite d'erreurs d'échantillonnage annuelles corrélées avec comme vecteur de

moyennes et matrice de covariances $(\mathbf{0}, \Sigma_a)$.

Se servant de ces modèles, Hillmer et Trabelsi obtiennent des estimations infra-annuelles étalonées en appliquant les moindres carrés stochastiques, c'est-à-dire en minimisant l'erreur quadratique moyenne, $E(\theta_t - \hat{\theta}_t)^2$. Dans la terminologie de l'analyse chronologique, cette méthode est aussi appelée extraction de signal et Hillmer et Trabelsi en font l'illustration dans leur article.

Les estimations infra-annuelles sont souvent entachées d'une erreur systématique à cause de problèmes de couverture. Le sous-dénombrement tient au fait que de nouvelles entreprises tardent à être intégrées à la base de sondage et que les entreprises qui n'ont pas d'employés (il s'agit le plus souvent de petites entreprises) ne figurent pas dans cette base. Les estimations infra-annuelles proviennent habituellement d'échantillons se chevauchants relativement petits, ce qui fait que les variances d'échantillonnage sont assez élevées et qu'il existe des covariances d'échantillonnage pour les estimations infra-annuelles qui se rapportent à des périodes différentes. En outre, la plupart des enquêtes économiques infra-annuelles produisent des séries d'estimations pour un certain nombre d'activités industrielles qui se déroulent dans un certain nombre de régions géographiques. Ces estimations sont publiées infra-annuellement dans des tableaux croisés (industrie \times région géographique) dont il faut étalonner les diverses composantes (fréquences par case, totaux marginaux et total général).

Pour ce qui a trait aux estimations annuelles, on peut supposer qu'elles sont sans biais puisqu'en pratique, les bases de sondage correspondantes sont relativement peu touchées par les problèmes de couverture. En outre, les estimations annuelles proviennent habituellement de recensements ou d'échantillons d'enquête assez grands, ce qui fait que les erreurs d'échantillonnage rattachées à ces estimations sont relativement faibles ou inexistantes alors que les covariances d'échantillonnage tendent à être élevées à cause du fort taux de chevauchement des échantillons d'une année à l'autre. Une autre chose qu'il faut retenir au sujet des estimations annuelles est que ces chiffres sont produits environ deux ans après la période de référence de l'enquête. Par exemple, les données annuelles pour 1988 n'ont été diffusées qu'en 1990 tandis que les données infra-annuelles sont diffusées habituellement quelques mois suivant la période à laquelle elles se rapportent. Ainsi, lorsque vient le moment d'étalonner les estimations infra-annuelles, il y a des périodes infra-annuelles pour lesquelles nous n'avons pas de repères annuels.

Une méthode d'étalonnage doit présenter un certain nombre de caractéristiques pour pouvoir être appliquée à des estimations de grandes enquêtes. Premièrement, elle doit être suffisamment simple pour qu'on puisse l'utiliser sans avoir à recourir trop largement à l'analyse de données. Deuxièmement, elle doit pouvoir produire des facteurs d'étalonnage préliminaire pour des périodes pour lesquelles il n'existe pas encore de données repères. Cette caractéristique permet de faire de l'étalonnage à mesure que sont produites les données infra-annuelles, sinon des discontinuités sont introduites dans ces données. Enfin, il est souhaitable qu'une méthode d'étalonnage assure la concordance des totaux généraux, des totaux marginaux et des estimations par case pour des données étalonnées.

Pour une analyse plus approfondie des deux dernières caractéristiques, voir Laniel et Fyfe (1989, 1990) et Cholette (1988a, 1988b). Dans le reste de cet article, nous voyons comment étalonner une seule série chronologique dans les conditions décrites ci-dessus.

3. ÉTALONNAGE D'UNE SÉRIE CHRONOLOGIQUE

Nous allons exposer ci-dessous quatre méthodes qui peuvent servir à étalonner une série d'estimations infra-annuelles sur les flux ou les stocks.

3.1 Méthode de Denton

Dans son article de 1971, Denton a proposé des méthodes d'étalonnage fondées sur la minimisation quadratique. Chacune de ces méthodes correspond à une fonction de perte particulière et l'une de ces fonctions s'exprime sous forme d'écarts proportionnels entre la série originale et la série étalonnée et est souvent utilisée pour résoudre le problème d'étalonnage exposé dans la section 2. Nous pouvons présenter cette méthode de Denton en termes statistiques en affirmant tout d'abord que les estimations infra-annuelles sont expliquées par le modèle:

Étalonnage des séries économiques

NORMAND LANIÉL et KIMBERLEY FYFE¹

RÉSUMÉ

L'étalonnage est l'opération qui consiste à améliorer les estimations tirées d'une enquête infra-annuelle à l'aide des estimations correspondantes tirées d'une enquête annuelle. Par exemple, les estimations de l'enquête annuelle sur les ventes au détail peuvent servir à améliorer les estimations des ventes au détail mensuelles. Dans cet article, nous nous penchons tout d'abord sur le problème que pose l'étalonnage des séries chronologiques issues d'enquêtes économiques et nous analysons les solutions les plus appropriées dans les circonstances. Dans un deuxième temps, nous proposons deux nouvelles méthodes statistiques qui reposent sur un modèle non linéaire pour données infra-annuelles. Finalement, nous obtenons des estimations étalonnées en appliquant la méthode des moindres carrés pondérés.

MOTS CLÉS: Erreurs d'enquête; modèle non linéaire; moindres carrés pondérés.

1. INTRODUCTION

L'étalonnage a toujours été défini comme l'opération qui consiste à ajuster des valeurs mensuelles ou trimestrielles tirées d'une source particulière à des valeurs annuelles (données repères) tirées d'une autre source (voir Denton 1971, Chollette 1988a et Monsour et Trager 1979). Il peut s'agir, par exemple, de corriger les chiffres des expéditions mensuelles des manufacturiers canadiens de manière que leur somme égale la valeur des expéditions établie à l'aide de l'enquête annuelle sur les manufactures. L'étalonnage peut aussi se définir, d'une manière plus générale, comme l'opération qui consiste à améliorer les estimations infra-annuelles tirées d'une source particulière à l'aide d'estimations annuelles provenant d'une autre source (voir Hillmer et Trabelsi, 1987). Contrairement à la première définition, la seconde suppose que les valeurs annuelles peuvent être erronées. Il peut s'agir, par exemple, d'améliorer les estimations des stocks mensuels des détaillants canadiens, établies à l'aide d'une enquête par sondage, au moyen des données sur les stocks de fin d'année tirées de l'enquête annuelle par échantillon sur le commerce de détail. La seconde définition est celle qui s'applique le plus souvent aux séries économiques et c'est à elle que nous intéressons dans cet article.

Cet article se divise en deux parties. Premièrement, nous allons exposer en détail le problème de l'étalonnage tel qu'il se présente pour de nombreuses séries chronologiques issues des grandes enquêtes économiques. Ensuite, nous allons présenter et analyser deux méthodes d'étalonnage parmi les plus courantes pouvant traiter une série à la fois. Comme aucune de ces méthodes ne répond parfaitement aux exigences, nous en proposons deux autres basées sur les moindres carrés pondérés avec modèle non-linéaire. Enfin, nous donnons un exemple de deux méthodes évoquées ci-dessus en nous servant de données simulées, puis nous en faisons l'analyse.

2. POSITION DU PROBLÈME

L'objet de cet article est de tenter d'améliorer des séries d'estimations infra-annuelles à l'aide de séries annuelles tirées des enquêtes-entreprises. Nous allons décrire ici les caractéristiques des données originales et ce que nous attendons d'une méthode d'étalonnage.

¹ Normand Lanier et Kimberley Fyfe, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario), K1A 0T6.

- CLIFF, A.D., et ORD, J.K. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society*, 37, 297-348.
- COHEN, A. (1983). Seasonal daily effect on the number of births in Israel. *Applied Statistics*, 32, 228-235.
- CRESSIE, N., et READ, T.R.C. (1989). Spatial data analysis of regional counts. *Biometrical Journal*, 31, 699-719.
- CROUCH, E.A.C., et SPIEGELMAN, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp \{ -t^2 \} dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.
- DEAN, C., LAWLESS, J.F., et WILLMOT, G.E. (1989). A mixed Poisson-inverse-Gaussian regression model. *La Revue Canadienne de Statistique*, 17, 171-182.
- DYN, N., et WAHBA, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal on Mathematical Analysis*, 13, 134-152.
- FRANKE, R. (1982). Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38, 181-200.
- GILCHRIST, W.G. (1967). Methods of estimation involving discounting. *Journal of the Royal Statistical Society*, 29, 355-369.
- HINDE, J. (1982). Compound regression models. GLIM 82 (éd. R. Gilchrist), 109-121. *Lecture Notes in Statistics*, 14. New York: Springer-Verlag.
- MALLOWS, C., et TUKEY, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (éds. Tiago de Oliveira, J. et coll.), 111-172. London: Academic.
- MANTON, K.G., WOODBURY, M.A., STALLARD, E., RIGGAN, W.B., CREASON, J.P., et PELOM, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates. *Journal of the American Statistical Association*, 84, 637-650.
- MIVAOKA, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. *La Revue Canadienne de Statistique*, 17, 123-140.
- PELTO, C.R., ELKINS, T.A., et BOYD, H.A. (1968). Automatic contouring of irregularly spaced data. *Geophysics*, 33, 424-430.
- PIERCE, D.A., et SANDS, B.R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Department, Oregon State University.
- PREPARATA, F.P., et SHAMOS, I. (1985). *Computational Geometry*. New York: Springer-Verlag.
- SHABAN, S.A. (1988). Poisson-lognormal distributions. *Lognormal Distributions*, 195-210 (éds. E.L. Crow et K. Shimizu). New York: Marcel Dekker.
- STANISWALIS, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.
- STONE, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595-620.
- TIBSHIRANI, R., et HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.
- TOBLER, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519-536.
- TSUTAKAWA, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83, 37-42.
- TUKEY, J.W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Dans The Collected Works of J.W. Tukey, Vol. 5 (1988), (éd. W.S. Cleveland). Pacific Grove: Wadsworth.
- TUKEY, J.W. (1990). Graphical displays of: Are the (x,y) pairs compatible with a linear dependence? Technical Report No. 301, Princeton University.

ANNEXE II

Pour plus de simplicité, considérons le cas d'un processus ponctuel $\{x_j\}$ avec une fonction de taux v sur la ligne. La fonction de vraisemblance logarithmique pondérée (localement) pour un processus de Poisson est, à une constante près,

$$\sum_j W(x - x_j) \log v(x_j) - \int W(x - u) v(u) du.$$

Par conséquent, l'estimateur (à pondération locale) du taux est

$$\hat{v}(x) = \frac{\sum_j W(x - x_j)}{\int W(x - u) du},$$

qui est la forme habituelle. Supposons maintenant que la ligne est découpée en intervalles R_i et que l'effectif connu est $N(R_i)$. On cherche à calculer

$$\sum_{x_j \in R_i} W(x - x_j).$$

Si on remplace l'expression ci-dessus par une mesure qui en est une approximation, $\Theta N(R_i)$, alors par la méthode des moments, on obtient

$$\Theta = \frac{\int_{R_i} W(x - u) du}{|R_i|}$$

ce qui nous amène à l'équation (3).

BIBLIOGRAPHIE

ABRAMOWITZ, M., et STEGUN, I.A. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington.

AITCHISON, J., et HO, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.

BOCK, R.D., et LIEBERMAN, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.

BRILLINGER, D.R. (1977). Discussion of Stone (1977). *The Annals of Statistics*, 5, 622-623.

BRILLINGER, D.R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42, 693-734.

BRILLINGER, D.R. (1990). Mapping aggregate birth data. Rapport technique, Statistics Department, University of California, Berkeley.

BRILLINGER, D.R., et PREISLER, H.K. (1983). Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*, 31-65. New York: Academic Press.

CLAYTON, D., et KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

CLEVELAND, W.S., et DEVLIN, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

CLEVELAND, W.S., et KLEINER, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics*, 17, 447-454.

Enfin, l'arbitre a exprimé des commentaires qui mettent très bien en relief les hypothèses et les limites de cette étude. Nous poursuivons notre recherche en vue de corriger les faiblesses qu'il signale. Nous avons jugé raisonnable ici de reproduire textuellement ses commentaires.

«Le choix des poids est trop *spécifique* et exige une plus grande réflexion. Si nous avons, par exemple, deux divisions de même superficie mais dont les populations respectives N_j sont très différentes, devons-nous leur attribuer le même poids? Tout dépend de ce qu'on juge le plus important entre la superficie ou la densité de population. En optant pour la seconde, nous pourrions faire abstraction du niveau de détail indu que l'on trouve dans la partie septentrionale de la province.»

«Les N_j posent des embûches que l'auteur semble ne pas ignorer; néanmoins, nous tenons à mettre en garde d'avantage le lecteur. Il serait bon d'avoir des mesures approximatives de la variance (dans la section 1, l'auteur indique qu'il n'en produit pas). On ne peut pas vraiment interpréter la figure 3 puisque les valeurs positives ou négatives peuvent être attribuables à des fluctuations aléatoires autour de zéro. Les contours de la figure 6 sont calculés avec un degré de précision très variable et ne sont pas comparables à certains égards. Enfin, pour ce qui a trait à l'estimation de α , β et γ (section 6), il serait tentant (mais imprudent) de supposer que ces valeurs sont significatives.»

«Toutes les variables aléatoires mentionnées dans cet article sont supposées indépendantes. Une autre façon de justifier les modèles pondérés est de supposer l'existence d'une distribution multidimensionnelle, où la moyenne conditionnelle à (x, y) , compte tenu des données relatives aux positions voisines, est une combinaison pondérée de ces données. On observe alors un lien de dépendance dans la distribution conjointe.»

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à G. Brackstone, R. Gussela, R. Raby, B. Sander, P. Spector, M. Subhani et R. Villani pour lui avoir procuré les données et les cartes et l'avoir assisté dans la géométrie computationnelle et le calcul en parallèle. John Tukey, Rob Tibshirani et l'arbitre ont exprimé des commentaires très utiles sur la première version. Cette étude a été rendue possible grâce à une subvention de la National Science Foundation (n° DMS-8900613).

ANNEXE I

Cette annexe contient quelques détails sur les calculs. Les limites de la province et des divisions de recensement forment des polygones. Pour calculer les poids $w_i(x, y)$, il a fallu utiliser un algorithme qui vérifiait si un point donné se trouvait dans un polygone donné. Pour calculer la moyenne et la variance d'un point aléatoire à l'intérieur d'un polygone donné, il a fallu recourir à un algorithme par lequel le polygone était divisé en triangles. Ces algorithmes sont analysés notamment dans Preparata et Shamos (1985). La fonction de vraisemblance approximative a été maximisée au moyen du programme FORTRAN va09a de Harwell. Pour le calcul en parallèle, nous avons décomposé la grille de 40 par 40 en 20 segments disjoints, puis avons effectué les calculs avec 20 postes de travail différents. Comme dans Brillinger et Preisler (1983), nous avons introduit des facteurs dans la fonction de vraisemblance afin de stabiliser les calculs. Miyaoka (1989) a observé que les calculs peuvent être sensibles au nombre de noeuds utilisés. Dans le cas qui nous occupe, nous avons augmenté le nombre de noeuds jusqu'à ce qu'il n'y ait plus de changement perceptible dans les résultats. Enfin, il y a aussi la question du choix des valeurs initiales. Dans cette étude, nous avons eu recours à la méthode des estimateurs de moments; toutefois, ceux-ci sont peut-être trop peu efficaces.

9. APPENDA

Dans cet article, nous avons notamment considéré l'inclusion d'un terme d'erreur (ϵ) dans le modèle afin de représenter les covariables pertinentes qui étaient absentes du modèle. Cela nous a amenés à utiliser la distribution normale logarithmique de Poisson. Tukey (1990) construit un "indice d'urbanisation" pour les divisions de recensement. Cet indice repose sur les chiffres de population des trois plus grandes agglomérations de la division. Les valeurs de l'indice, x_i , sont reproduites dans la figure 14; nous constatons que les valeurs les moins élevées correspondent aux divisions de recensement qui renferment les villes de Regina et de Saskatoon.

Le tableau ci-dessous contient les résultats de l'ajustement de modèles de Poisson pour B_{ijk} étant donné N_i , au moyen du logiciel GLIM; ces modèles sont: i) $N_i \exp\{\alpha + \beta_j + \gamma_k\}$, ii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta x_i\}$ et iii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta_1 x_i + \delta_2 x_i^2\}$.

Variables	Somme des carrés des écarts	d.l.	valeur p
jours ouvrables, année	227.3	69	
+ urbanisation	86.69	68	
+ urbanisation**2	83.13	67	.088

Grâce à l'introduction de la variable d'urbanisation, x_i , nous pouvons maintenant nous accommoder d'un modèle de Poisson ordinaire.

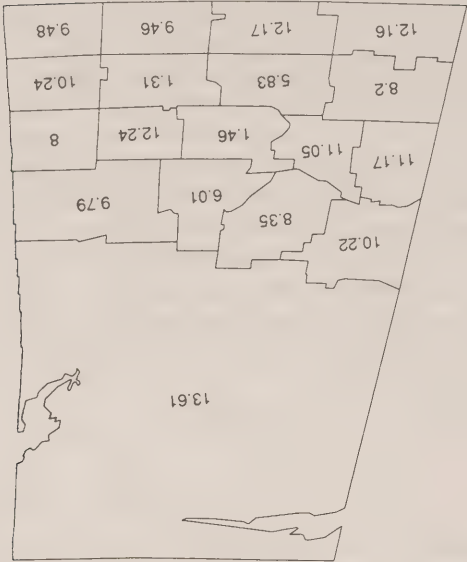


Figure 14. Valeurs de l'indice d'urbanisation de Tukey.

Toutes les études qui traitent l'estimation par le modèle normal logarithmique de Poisson et dont nous connaissons l'existence impliquent une forme quelconque d'approximation. Clayton et Kaldor (1987), par exemple, utilisent une équation quadratique en guise d'approximation pour la fonction de vraisemblance logarithmique conditionnelle de Poisson, et Aitchison et Ho (1989) utilisent à leur tour l'intégration numérique, non sans avoir transformé les paramètres au préalable. Crouch et Spiegelman (1990) ont proposé récemment une nouvelle forme d'approximation, dont on n'a pas encore étudié l'efficacité par rapport au modèle normal logarithmique de Poisson.

8. ANALYSE

L'analyse à pondération locale et les modèles avec effets aléatoires semblent être deux moyens de résoudre avec souplesse toute une série de problèmes ayant trait à des données régionales. Les termes d'effet aléatoire ont deux fonctions importantes: tenir compte des effets manquants et "emprunter" de l'information de manière à produire de meilleures estimations pour les principaux paramètres. En ce qui concerne le modèle de Poisson ordinaire, les totaux élémentaires sont efficaces mais, dans le cas qui nous occupe, il existe une variation non représentée par la loi de Poisson à cause des variables manquantes.

La méthode exige beaucoup de temps d'ordinateur parce que l'intégration numérique et l'estimation par le maximum de vraisemblance sont exécutées à de nombreuses positions sur une grille; toutefois, les opérations se sont bien déroulées sur le système Sun 3/50 de Berkeley.

Beaucoup de recherches restent à faire; mentionnons au passage des sujets comme l'évaluation de l'ajustement, le calcul et la présentation de la variabilité, le choix de la fonction de poids (notamment le choix de τ dans l'équation (4)), les analyses pour d'autres groupes d'âge et d'autres provinces et la définition d'asymptotes appropriées. Il faudra aussi chercher à combi- prendre pourquoi, avec des valeurs initiales proches de la valeur réelle, le processus d'optimi- sation convergeait parfois vers des valeurs estimées assez éloignées de la valeur réelle. L'avantage de notre étude est qu'elle renferme une quantité d'autres données qui peuvent servir au fur et à mesure des recherches. En examinant la figure 6 et les suivantes, on remarque une faiblesse importante de la technique: un trop grand niveau de détail dans la partie septentrionale de la province.

Parmi les articles récents portant sur l'analyse des données démographiques, notons ceux de Cressie et Read (1989), de Clayton et Kaldor (1987), de Tsutakawa (1988) et de Manton et coll. (1989). Contrairement à la présente étude, ces articles n'ont pas pour objet d'étudier la question des surfaces lisses.

Il est amusant de savoir que l'existence d'un phénomène d'une période de sept jours nous a amenés à découvrir très tôt qu'il y avait eu confusion au sujet de la série de données à utiliser. En effet, lorsqu'est venu le moment de déterminer, à l'aide de la série initiale de données, les jours où il y avait le moins de naissances, nous avons constaté qu'il s'agissait, selon toute vraisemblance, du vendredi et du samedi parce que nous avions en mains les données de 1987 et non celles de 1986, comme ç'aurait dû être le cas.

Une fois les analyses terminées, nous avons appris que le nombre de naissances avait été établi en fonction des divisions de recensement de 1981 tandis que les chiffres de population, eux, l'avaient été en fonction des divisions de 1986. Heureusement, les limites des divisions ont peu changé entre les deux recensements mais cette méprise offre une raison de plus pour réclamer une méthode qui puisse tenir compte de la variation non représentée par une loi donnée.

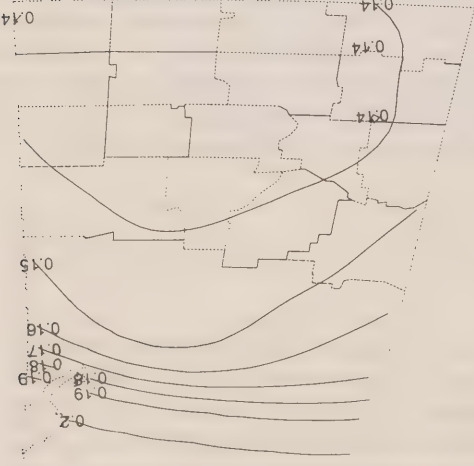


Figure 11. Tracé comparable à celui de la figure 8, sauf que (comme dans la figure 10) un terme d'erreur normal a été ajouté au prédicteur linéaire.

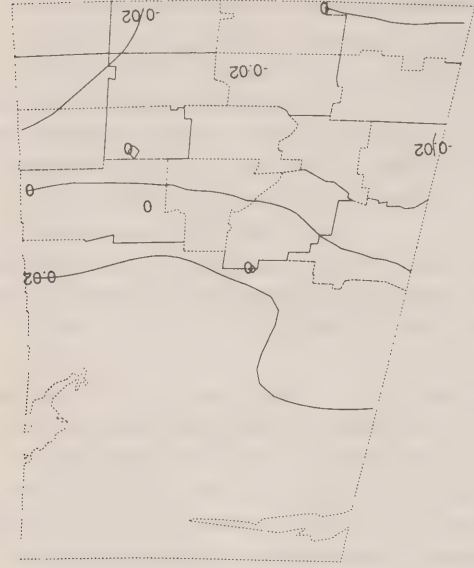


Figure 13. Erreur type estimée, $\hat{\sigma}(x,y)$, du terme normal qui a été ajouté au prédicteur linéaire.

7. AJUSTEMENT DU MODELE NORMAL LOGARITHMIQUE DE POISSON

Etant donné une variable explicative multidimensionnelle x , un modèle de Poisson de moyenne $N \exp\{x\theta\}$ pour B pourrait expliquer assez bien les données. Comme variables explicatives, pensons au régime alimentaire, au mode de vie, aux conditions atmosphériques, à l'environnement, aux jours fériés, au changement démographique, à la structure par âge, aux caprices des délimitations. Comme nous ne connaissons rien de ces variables dans les circonstances, nous allons supposer que les variables absentes du modèle sont représentées collectivement par une variable d'erreur. Nous allons aussi supposer que, étant donné ϵ , la variable aléatoire B suit une distribution de Poisson de moyenne $N\mu \exp\{\epsilon\}$ et que ϵ suit une distribution normale de moyenne 0 et de variance σ^2 . En ce qui concerne ce modèle, on dira que B suit une distribution normale logarithmique de Poisson. Pour plus de renseignements sur cette distribution, veuillez vous référer à Shaban (1988). Il arrive que ϵ découle directement du contexte (voir Brillinger et Preisler (1983) pour un exemple) mais dans le cas qui nous occupe, nous avons simplement supposé son existence.

Un inconvénient majeur du modèle logarithmique de Poisson est qu'il n'existe pas d'expression en forme analytique pour la fonction de probabilité. Cependant, le modèle se prête très bien à l'introduction d'effets et au traitement de variables manquantes. Les travaux de Bock et Lieberman (1970), de Pierce et Sands (1975) et de Hinde (1982) nous indiquent que l'on peut recourir à la quadrature numérique. Nous pouvons exprimer la fonction de proba-

bilité par la formule

$$p(X) = \frac{1}{I} \int (ve^{\alpha z})^X \exp\{-ve^{\alpha z}\} \phi(z) dz$$

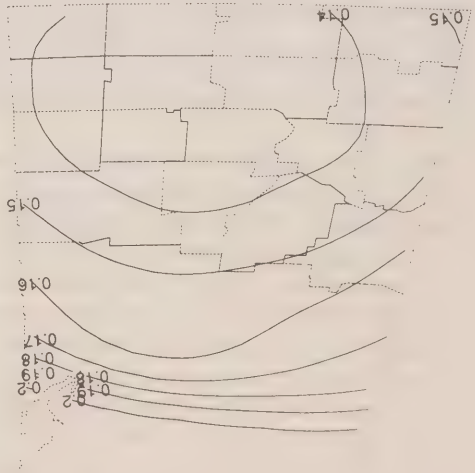
où ϕ est la densité normale standard, X correspond à B et v correspond à $N\mu$. Pour faire une analyse de données, on remplace l'intégrale par un nombre fini de termes comprenant des noeuds, z_i , et des poids, w_i ,

$$p(X) \approx \frac{1}{L} \sum_{i=1}^L Y_i (ve^{\alpha z_i})^X \exp\{-ve^{\alpha z_i}\} w_i.$$

Pour une liste de noeuds et de poids, voir, par exemple, Abramowitz et Stegun (1964). Les figures 10, 11, 12 et 13 donnent les résultats de l'ajustement du modèle normal logarithmique de Poisson (y compris les versions avec effet des jours ouvrables et effet d'année) fait à l'aide de $L = 5$ noeuds. Le modèle suppose que B_{ijk} , étant donné N_j et Z , suit une distribution de Poisson de moyenne

$$N_j \exp\{\alpha + \beta_j + \gamma_k + \sigma Z\}$$

Z étant un écart normal standard, et suppose de plus que les différents Z sont indépendants. Dans ce modèle, i désigne la division de recensement, j indique s'il s'agit d'un jour de semaine ou non et k désigne l'année. La figure 10, qui représente un tracé de contours pour $\exp\{\alpha(x,y)\}$, montre des cercles vaguement concentriques autour des régions urbaines, comme dans la figure 7. La forme irrégulière du tracé donne à penser que le processus d'estimation pourrait, à une occasion, avoir convergé vers un extrémum local. Les figures 11 et 12 donnent les tracés de contours pour $\beta(x,y)$ et $\gamma(x,y)$ respectivement. Là encore, on a l'impression qu'il y a eu convergence vers un extrémum local. La figure 13, qui représente un tracé de contours pour $\hat{\sigma}(x,y)$, n'est pas facile à décrire. Elle donne à croire que la valeur estimée $\hat{\sigma}$ est passablement variable. Cette valeur se situe autour de 0.1 et est donc comparable à l'effet des jours ouvrables défini dans la section 6.



6. AJUSTEMENT DU MODÈLE DE POISSON

Dans cette analyse, nous nous intéressons uniquement aux femmes de 25 à 29 ans et aux naissances enregistrées chez ce groupe de femmes. Désignons par $i = 1, \dots, 18$ la division de recensement et par N_i le nombre de femmes recensées dans la division i (il s'agit des chiffres du recensement du 3 juin 1986). Soit B_i le nombre total de naissances enregistrées chez les femmes de 25 à 29 ans en 1986 et 1987.

Supposons que la distribution de probabilité $p(\cdot)$ de B_i (section 5) est une distribution de Poisson de moyenne $2N_i\mu$ (le chiffre 2 signifiant que le paramètre μ représente un taux annuel de natalité). Cette hypothèse repose sur l'idée que les anniversaires de naissance sont aléatoires (voir Brillinger 1986).

Étant donné l'hypothèse du modèle de Poisson, la valeur estimée (à pondération locale) du taux annuel de natalité à la position (x,y) est

$$\hat{\mu}(x,y) = \sum_i^i w_i(x,y) B_i / 2 \sum_i^i w_i(x,y) N_i. \tag{5}$$

Ces valeurs sont calculées pour (x,y) sur une grille de 40 par 40 et le tracé de contours correspondant est reproduit dans la figure 6. On observe une progression lente de la valeur des contours. Le taux varie de .14 à .20, les valeurs les plus élevées étant observées dans la partie septentrionale de la province et les moins élevées étant concentrées dans les régions les plus urbanisées.

Nous avons mentionné plus haut que les données à l'étude avaient une dimension temporelle importante. Les modèles doivent tenir compte de cette réalité. En particulier, ils doivent pouvoir décrire la période hebdomadaire qui caractérise ces données, ainsi que les tendances démographiques possibles. Voici donc un modèle qui mérite d'être considéré. Soit j une variable indicatrice qui prend la valeur 1 si c'est pour un jour de semaine et la valeur 2 si c'est pour un jour non ouvrable. Soit k une seconde variable indicatrice qui prend la valeur 1 pour 1986 et la valeur 2 pour 1987. Désignons par B_{ijk} le nombre de naissances dans la division de recensement i . Supposons que B_{ijk} étant donné N_i , suit une distribution de Poisson de moyenne $N_i \exp\{\alpha + \beta_j + \gamma_k\}$. β_j représente l'effet des jours ouvrables et γ_k , l'effet d'année, et nous supposons que $\beta_1 + \beta_2, \gamma_1 + \gamma_2 = 0$ pour que le modèle soit identifiable. S'il n'y a pas d'effet des jours ouvrables, alors, $\beta_1, \beta_2 = 0$. S'il n'y a pas d'effet d'année, alors $\gamma_1, \gamma_2 = 0$. Par ailleurs, en utilisant la méthode d'estimation à pondération locale décrite dans la section 5, on peut estimer α, β et γ et comme des fonctions de la position (x,y) . (Pour simplifier les calculs, nous avons utilisé uniquement les 364 ($= 7 \times 52$) premiers jours de chaque année.)

La figure 7 donne la valeur estimée $\exp\{\hat{\alpha}(x,y)\}$ du taux annuel de natalité. Il est intéressant de constater que, par rapport à la figure précédente (modèle de Poisson ordinaire), les contours sont un peu plus éloignés des régions urbaines. La figure 8 illustre l'effet estimé des jours ouvrables, $\hat{\beta}_1(x,y)$. Dans ce cas, on observe une poussée dans l'est de la province. Cette figure tranche considérablement avec la figure 4, où l'on se contente de reproduire de simples écarts par région. Le trait distinctif de la figure 8 est qu'elle reflète les différences de population. Les valeurs de $\hat{\beta}$ vont de .08 à .13 tandis que celles de $\hat{\alpha}$ vont de -2.1 à -1.6. La figure 9 donne l'effet d'année estimé, $\hat{\gamma}_1(x,y)$. Les valeurs de cet effet vont de -.03 à .03. En termes numériques, l'effet des jours ouvrables est plus grand que l'effet d'année.

L'analyse que nous venons de faire donne à penser que des variables fondamentales peuvent influencer sur les taux de natalité et que nous devons en tenir compte dans la modélisation et l'analyse.

Tibshirani et Hastie (1987) ont élaboré une méthode d'estimation à équipondération locale fondée sur le principe de la vraisemblance. Cleveland et Devlin (1988) font un traitement très détaillé de la méthode des moindres carrés. Pour sa part, Staniswalis (1989) étudie et applique le cas p général. Les avantages de la technique à pondération locale sont les suivants: aucune hypothèse "cachée" sur la distribution du modèle, mise en évidence des cas de non-additivité, variantes pour la résistance et l'influence, additivité simple des observations et aucune inversion de matrice (comme l'exige le krigeage par exemple). Les données qui nous intéressent ici consistent essentiellement en des totaux pour des divisions de recensement. Nous ne pouvons donc pas utiliser directement la méthode décrite dans la section précédente. Il s'agit ici de déterminer des poids $w_i(x,y)$ qui traduisent convenablement l'effet de la division de recensement i sur la position (x,y) . Supposons que $|R_i|$ désigne la superficie de la division de recensement i . Alors, la fonction de poids élémentaire est

$$w_i(x,y) = \frac{1}{|R_i|} \text{ pour } (x,y) \text{ dans } R_i$$

et 0 dans le cas contraire. Dans cette étude, nous allons utiliser des fonctions fondamentales comme

$$w_i(x,y) = \frac{1}{|R_i|} \int_{R_i} W(x - u, y - v) du dv \quad (3)$$

où $W(\cdot)$ est un noyau qui convient à des données non agrégées comme celles analysées, par exemple, dans Cleveland et Devlin (1988). Pour justifier l'utilisation de l'équation (3), nous pouvons considérer un processus ponctuel de Poisson (voir Annexe II). Les estimations seront calculées à l'aide des formules (1) ou (2), W_i étant remplacé par w_i .

Dans cette étude initiale, les poids particuliers utilisés à $\mathbf{r} = (x,y)$ sont

$$w_i(\mathbf{r}) = \exp \{ -(1 - \rho)^2 \|\mathbf{r} - \mathbf{r}_i\|^2 / 2\tau^2 \} \quad (4)$$

à l'extérieur de l'ellipse $(\mathbf{r}_0 - \mathbf{r}_i)' S_i^{-1} (\mathbf{r}_0 - \mathbf{r}_i) = d_0^2 = 5.991$ et 1 à l'intérieur. Dans l'équation (4), $\|\mathbf{r}\|^2 = x^2 + y^2$, $\rho = d_0 / \sqrt{(\mathbf{r} - \mathbf{r}_i)' S_i^{-1} (\mathbf{r} - \mathbf{r}_i)}$ et $\tau = 0.025$, où $\mathbf{r}_i = E U_i$ et $S_i = \text{var } U_i$, U_i étant une variable aléatoire distribuée uniformément dans R_i . Le paramètre ρ est défini de manière que la fonction de poids soit continue. En clair, cela signifie que les divisions de recensement sont représentées par des ellipses ayant la même moyenne et la même matrice de variances-covariances. (Les valeurs exactes ont été déterminées après quelques essais pour faire en sorte que la surface de la première ellipse équivalente à environ 95% de la superficie de la division de recensement.) Nous aurions pu choisir d'autres figures que l'ellipse - un rectangle par exemple - mais la recherche sur ce sujet est encore jeune et il est permis de croire que les études ultérieures utiliseront des poids comme celui défini en (3).

La figure 5 illustre les contours à .50 et à .99 des poids $w_i(x,y)$ pour plusieurs divisions de recensement. On remarque que ces contours suivent la forme générale des divisions. La forme irrégulière de certains contours est attribuable au caractère discret de la grille de 40×40 utilisée dans les calculs.

Tobler (1979) et Dyn et Wahba (1982) décrivent d'autres fonctions de poids construites dans des circonstances semblables. La méthode que nous venons d'exposer présente quelques-uns des avantages de la technique à pondération locale: additivité des termes des expressions (1) et (2) et absence d'interaction; aucune inversion de matrice requise; et intégration facile d'un mécanisme de résistance aux valeurs aberrantes.

Cliff et Ord (1975) (section 5.1) cherchent à mesurer l'influence que des comtés exercent l'un sur l'autre. Notre article a plutôt pour but d'analyser l'influence d'un "comté" sur un point géographique en particulier. Peut-être le poids, qui rend compte de l'influence, devrait-il dépendre de covariables (par ex.: la population de comté)?

Il s'agit d'estimer Θ en maximisant la fonction de vraisemblance logarithmique pondérée

(1)

$$\sum_{i=1}^I W_i(x,y) \log p(Y_i | \Theta)$$

ou (ce qui revient souvent au même) en résolvant le système d'équations d'estimation

(2)

$$\sum_{i=1}^I W_i(x,y) \Psi(Y_i | \Theta) = 0$$

où $\Psi(Y | \Theta) = \partial \log p / \partial \Theta$, la fonction de caractérisation.
Afin d'illustrer la technique, prenons un cas simple, celui où Y suit une distribution normale de moyenne μ et de variance σ^2 . On obtient la valeur estimée (à pondération locale) de μ à (x,y) en minimisant l'expression

$$\sum_{i=1}^I W_i(x,y) [Y_i - \mu]^2$$

ce qui donne

$$\hat{\mu}(x,y) = \frac{\sum_{i=1}^I W_i(x,y) Y_i}{\sum_{i=1}^I W_i(x,y)},$$

expression à première vue intéressante. Soulignons que ces formules sont souvent utilisées en infographie pour l'interpolation de données (voir, par exemple, Franke (1982)).

Parmi les textes traitant ce sujet, mentionnons l'article de Gilchrist (1967) sur l'"actualisation", celui de Peltó et coll. (1968) sur les moindres carrés, celui de Cleveland et Kleiner (1975), qui propose l'utilisation de moyennes mobiles, et celui de Stone (1977), qui porte plus particulièrement sur la régression. Dans son analyse de l'article de Stone, Brillinger (1977) propose l'équation (2) pour une distribution générale et justifie sa proposition en assimilant cette équation à une règle de Bayes. Considérons, en particulier, la fonction de perte

$$L(Y | \Theta) = -\log p(Y | \Theta).$$

Supposons que nous voulons obtenir la valeur estimée d'un paramètre à $\mathbf{r} = (x,y)$. On peut exprimer le risque de Bayes par l'équation

$$E\{L(Y | \Theta_r)\} = E\{E\{L(Y | \Theta_r) | \mathbf{r}\}\}.$$

La règle de Bayes vise à déterminer

$$\min_{\Theta} E\{L(Y | \Theta) | \mathbf{r}\}.$$

A l'aide des données Y_i , des coordonnées \mathbf{r}_i et de $W_i(\mathbf{r})$, un noyau centré sur \mathbf{r}_i , on peut déterminer approximativement l'espérance conditionnelle par la formule

$$E\{\log p(Y | \Theta) | \mathbf{r}\} \approx \sum_{i=1}^I W_i(\mathbf{r}) \log p(Y_i | \Theta)$$

ce qui nous renvoie à l'expression (1).

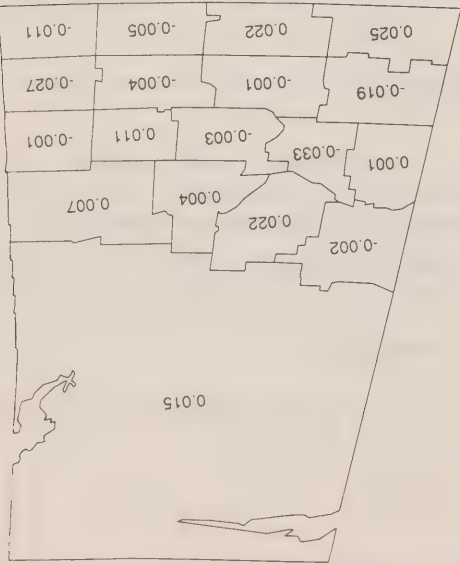


Figure 3. Différence entre le taux pour 1986 (mêmes données que pour la figure 2).

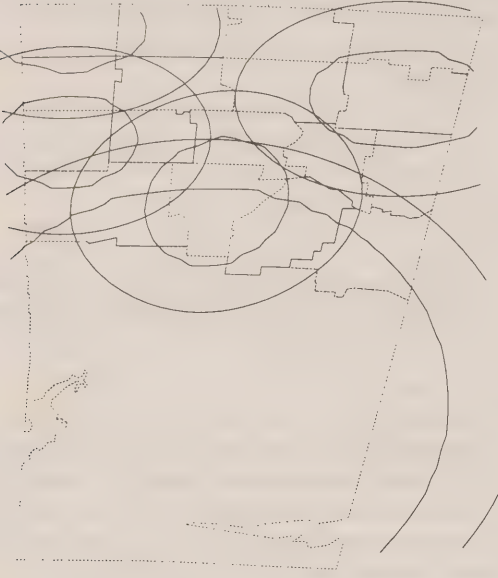


Figure 5. Poids $W_i(x,y)$, utilisés dans les équations (1) ou (2) et calculés au moyen de l'équation (4), pour quatre divisions de recensement. Les divisions ne sont pas toutes représentées pour des raisons de clarté. Les contours à .50 et .99 y sont reproduits.

semaine (jours de semaine = lundi au vendredi). Le premier est plus élevé que le second dans toutes les divisions sauf une. Ces résultats sont conformes à ceux de diverses autres études et, comme nous le disions dans la section précédente, s'expliquent fort probablement par le fait que les médecins agissent de manière à favoriser les accouchements les jours de semaine (pour ne pas devoir assister à des accouchements les fins de semaine).

Les taux indiqués dans les figures 2, 3 et 4 représentent des valeurs moyennes pour les divisions de recensement.

4. CONTRAINTES DE LA REPRÉSENTATION GÉOGRAPHIQUE DES DONNÉES

Méfiez-vous des cartes qui assignent une valeur moyenne à tout un territoire... (TRADUCTION)

Par ces mots, Tukey (1979) déplore l'utilisation de cartes comme celles représentées par les figures 2, 3 et 4, où une valeur unique est associée à chaque division géographique. En fait, en examinant la figure 2, il est plus logique de penser que le taux de natalité ne varie pas aussi brusquement d'une division de recensement à l'autre. Un des objectifs de notre étude est justement d'établir des cartes en courbes de niveau qui décrivent une variation progressive des taux de natalité. Nous espérons que ces cartes permettront d'en arriver à des descriptions stochastiques générales du phénomène et favoriseront des analyses exploratoires éclairantes.

Dans notre étude, nous nous intéressons aussi à la distribution statistique des effets proprement dits. Un modèle stochastique spécial qui s'impose naturellement dans les circonstances est le modèle de Poisson. Or, nous savons, d'après des études antérieures, que les naissances sont liées à de nombreuses variables socio-économiques comme le régime alimentaire, le mode de vie, les conditions atmosphériques, l'environnement, les jours de semaine, les jours fériés, la structure par âge. En outre, la population des divisions de recensement a fluctué autour des chiffres du recensement tout le long de l'année 1986 et 1987 et, finalement, l'âge des femmes visées par l'étude varie de 25 à 29 ans. En résumé, il semble que nous devions utiliser un modèle plus souple que celui de Poisson, c'est-à-dire un modèle qui pourrait tenir compte des covariables absentes. Nous utiliserons donc la distribution normale logarithmique de Poisson dans les circonstances. Par ailleurs, à cause de la présence du paramètre d'écart-type dans cette distribution, nous recourrons à l'"emprunt d'information" en combinant les valeurs d'observations de la manière décrite par Mallows et Tukey (1982). (Nous utilisons le terme "emprunt d'information" plutôt que "estimation empirique de Bayes", que certains préfèrent, parce qu'il a un sens plus large et qu'il est en usage depuis longtemps.) Dean et coll. (1989) est une autre source récente qui traite de la variation non représentée par une loi donnée.

5. ANALYSE À PONDERATION LOCALE

Dans le cas de données non agrégées, l'ajustement à pondération locale est une méthode appropriée pour estimer des quantités qui varient graduellement. Supposons que nous avons une variable aléatoire X avec une distribution de probabilité $p(X|\Theta)$ qui dépend du paramètre de dimension finie Θ . Supposons par ailleurs que nous voulons estimer Θ pour la position définie par les coordonnées (x_i, y_i) . Supposons enfin que nous connaissons X_i pour la position (x_i, y_i) . Nous allons définir un poids $W_i(x, y)$ qui dépend de la distance entre (x_i, y_i) et (x, y) .

nous définissons des modèles avec effets aléatoires, que nous ajustons en vue de traiter la variation non représentée par la loi de Poisson. Le second volet peut être envisagé comme un exercice visant à évaluer la capacité du modèle normal logarithmique de Poisson d'intégrer les covariables non mesurées et les erreurs. L'analyse à pondération locale consiste à construire des poids, $w_i(x,y)$, qui sont censés traduire l'effet de la division de recensement i (un agrégat) sur le lieu géographique défini par les coordonnées (x,y) . Les données de la division de recensement étant connues, on applique ensuite ces poids à chacun des termes de la fonction de vraisemblance logarithmique ou des équations d'estimation correspondantes, puis on calcule les valeurs estimées des paramètres.

Nous tenons à préciser qu'il s'agit là d'un rapport provisoire sur des recherches en cours. Par exemple, nous ne prenons pas en considération la structure très détaillée des données et ne produisons aucune mesure de la variance des diverses estimations. Les expressions que nous utilisons pour représenter les poids sont élémentaires et devraient subir des transformations au fur et à mesure des recherches; malgré cela, on peut penser que notre analyse conserve un certain intérêt par son caractère.

Dans un autre article portant sur le même sujet (Brillinger 1990), nous examinons la question sous l'aspect géographique seulement.

2. REPRÉSENTATION CHRONOLOGIQUE DU NOMBRE DE NAISSANCES

Le graphique de la partie supérieure de la figure 1 reproduit le nombre total de naissances enregistrées en Saskatchewan pour chaque jour de 1986. La ligne continue est le résultat d'un lissage accentué de la moyenne de naissances pour 1986. La ligne continue est le résultat d'un lissage accentué de la série et est censée faire ressortir toute tendance. Un examen sommaire de ce graphique ne nous révèle pas de phénomène particulier; cependant, lorsqu'on trace le périodogramme de la racine carrée du nombre de naissances (voir partie inférieure de la figure 1), on constate quelque chose d'intéressant. (La racine carrée sert à rendre la série plus symétrique et plus normale.) Les lignes continues du haut et du bas représentent des limites de confiance marginales à 95% environ pour une série fortement lissée. Un sommet est facilement observable à la fréquence .143 cycle/jour, ce qui correspond à une période de 7 jours. Ce phénomène périodique est fort connu dans l'analyse des données sur les naissances (voir, par exemple, Cohen (1983) et Miyaoaka (1989) et les bibliographies qu'ils contiennent). Il s'explique habituellement par le fait que les médecins interviennent et provoquent l'accouchement, surtout les jours de semaine.

3. REPRÉSENTATION GÉOGRAPHIQUE DU NOMBRE DE NAISSANCES

La figure 2 indique, pour chaque division de recensement, le taux annuel de natalité chez les femmes de 25 à 29 ans pour les années 1986 et 1987 combinées. On observe le taux annuel le plus élevé (.208 naissances par femme) dans la partie septentrionale de la province et les deux taux les plus faibles, dans les divisions de recensement où se trouvent les villes de Regina et de Saskatoon.

La figure 3 donne, pour chacune des 18 divisions de recensement, l'écart numérique entre les taux annuels de natalité pour 1987 et 1986. (Précisons que le chiffre de population de 1986 a servi de diviseur dans chaque cas.) Les écarts se situent autour de 0. Il convient toutefois de préciser que ces taux reposent sur des tailles de population très variées.

Dans la section précédente, nous avons constaté l'existence d'un phénomène ayant une période de 7 jours. La figure 4 indique, pour chaque division de recensement, la différence entre le taux de natalité moyen pour les jours de semaine et le taux moyen pour les fins de

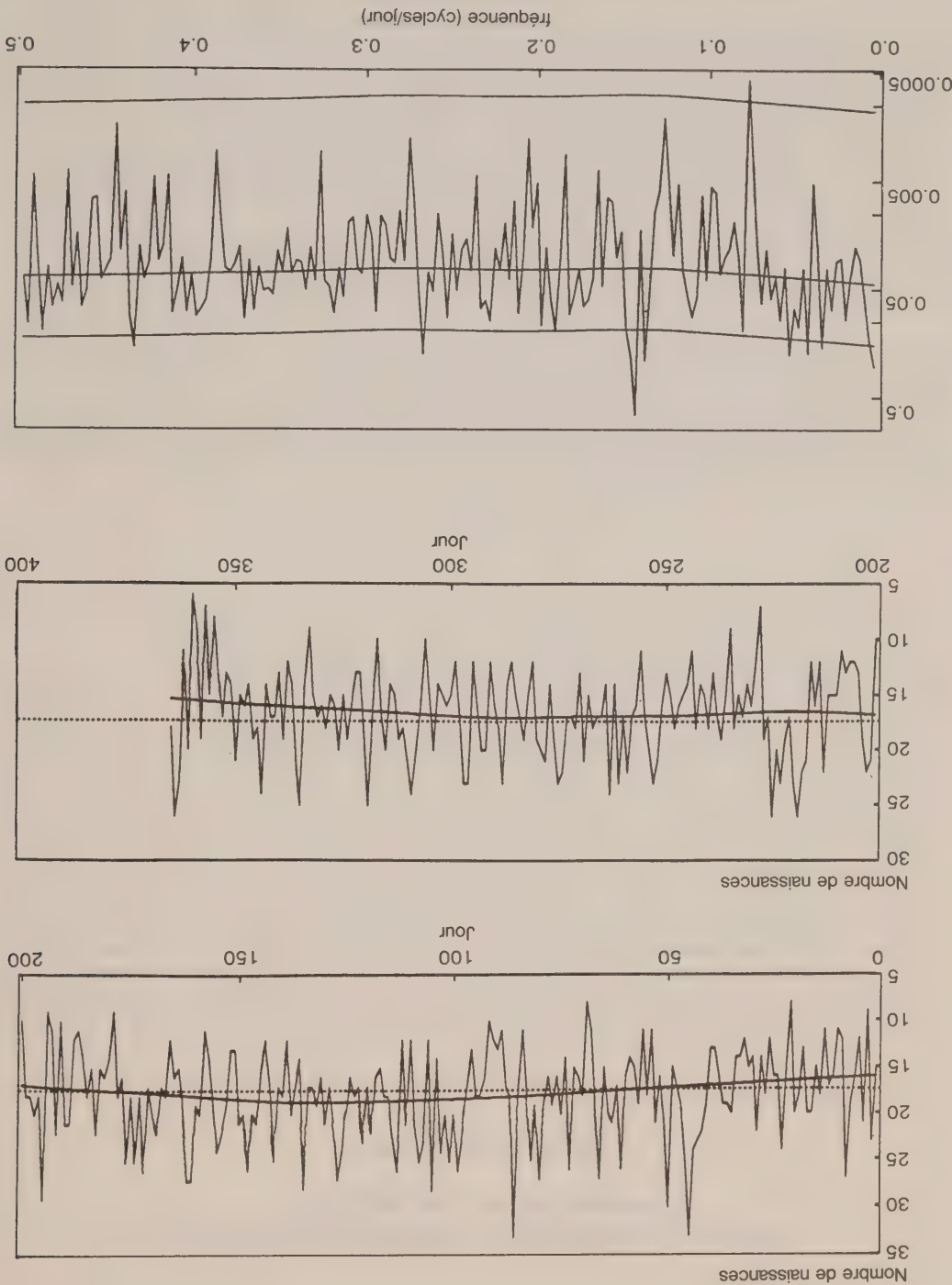


Figure 1. Partie supérieure: Série des naissances enregistrées chez les femmes de 25 à 29 ans en 1986 pour la province de Saskatchewan. Partie inférieure: Périodogramme de la racine carrée du nombre de naissances représenté dans le graphique de la partie supérieure. Les lignes continues représentent des limites de confiance à 95% environ. Le sommet correspond à une période de 7 jours.

Modélisation spatiale et temporelle de données agrégées sur les naissances

DAVID R. BRILLINGER¹

RÉSUMÉ

À l'aide de graphiques et de cartes de la province de Saskatchewan, nous faisons une analyse des naissances enregistrées en 1986 et 1987 par division de recensement. Nous cherchons à déterminer de quelle manière le nombre des naissances est lié aux périodes de l'année et aux régions géographiques; à cette fin, nous établissons des cartes en courbes de niveau qui décrivent le phénomène des naissances de façon uniforme. Le fait qu'il s'agit de données agrégées pose un problème majeur. En deuxième lieu, nous voulons vérifier dans quelle mesure le modèle normal logarithmique de Poisson peut remplacer, pour des données discrètes, le modèle de régression normal pour variables aléatoires continues. À cette fin, une hiérarchie de modèles pour variables aléatoires discrètes sont ajustés aux observations par la méthode du maximum de vraisemblance; il s'agit du modèle de Poisson ordinaire, du modèle de Poisson avec effet des années et des jours ouvrables et du modèle normal logarithmique de Poisson avec effet des années et des jours ouvrables, l'utilisation de ce dernier étant justifiée par l'absence de covariables importantes dans le processus d'ajustement. Comme nous l'indiquons dans l'article, il s'agit là de résultats provisoires.

MOTS CLÉS: Données agrégées; emprunt d'information; établissement de cartes en courbes de niveau; variation non représentée par la loi de Poisson; analyse à pondération locale; cartes; périodogramme; distribution de Poisson; distribution normale logarithmique de Poisson; effets aléatoires; données géographiques; séries chronologiques; covariables non mesurées.

1. INTRODUCTION

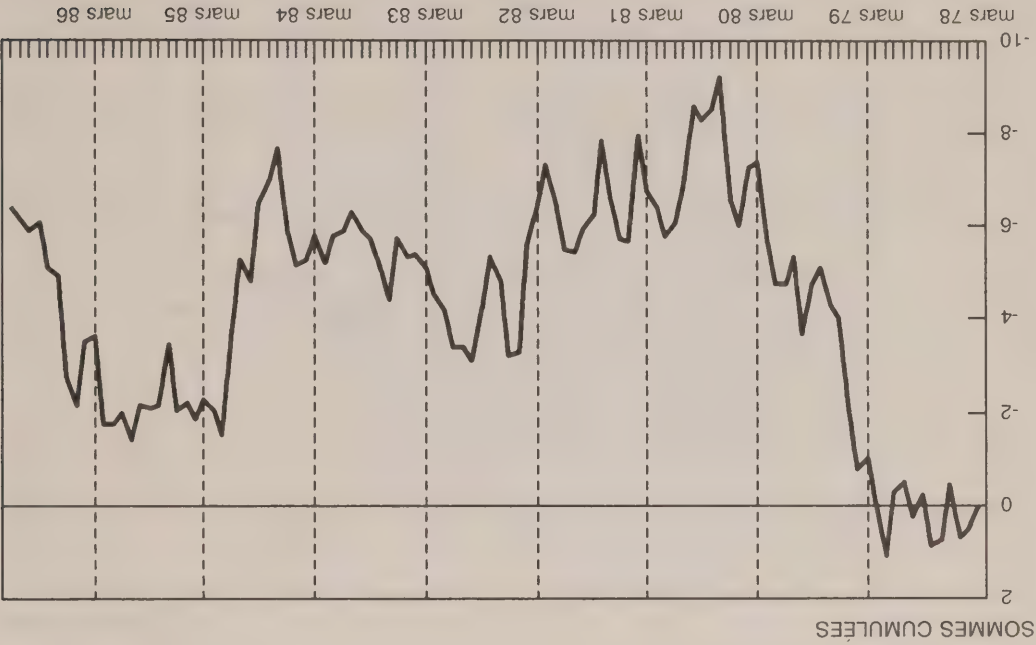
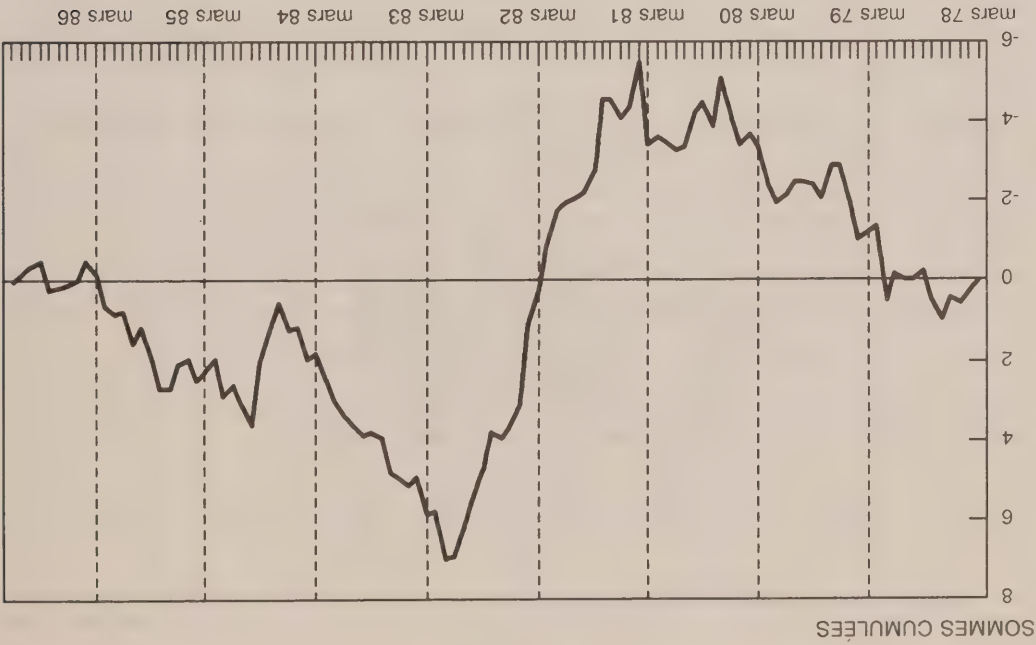
Dans cet article, nous nous intéressons à des données qui ont été enregistrées suivant des périodes de temps et des régions géographiques. Il devrait être facile d'analyser de telles données à cause des possibilités de représentation graphique (par exemple, taux en fonction de la période ou nombre en fonction de la région géographique, comme dans le cas de la représentation graphique de résidus, si souvent utilisée dans l'analyse de régression), dans le cas qui nous occupe toutefois, l'aggrégation d'éléments de base soulève des difficultés majeures. Les données analysées ici concernent essentiellement le nombre quotidien de naissances chez les femmes de 25 à 29 ans pour les années civiles 1986 et 1987 pour chacune des 18 divisions de recensement de la Saskatchewan. Nous nous servons aussi des chiffres de population correspondants, établis lors du recensement de 1986, pour le calcul de taux. Nous avons choisi la Saskatchewan pour cette étude pilote parce que son territoire est modérément étendu et que les limites de ce territoire et des divisions de recensement sont assez régulières. (La seconde raison était importante au début de l'étude parce que nous ne disposions pas de cartes produites par ordinateur.) Nous avons choisi les femmes de 25 à 29 ans parce que c'est le groupe d'âge auquel correspond le plus grand nombre de naissances. Les données nous ont été fournies par Statistique Canada. Elles ont pour caractéristique d'être agrégées, non gaussiennes et non stationnaires dans l'espace et le temps.

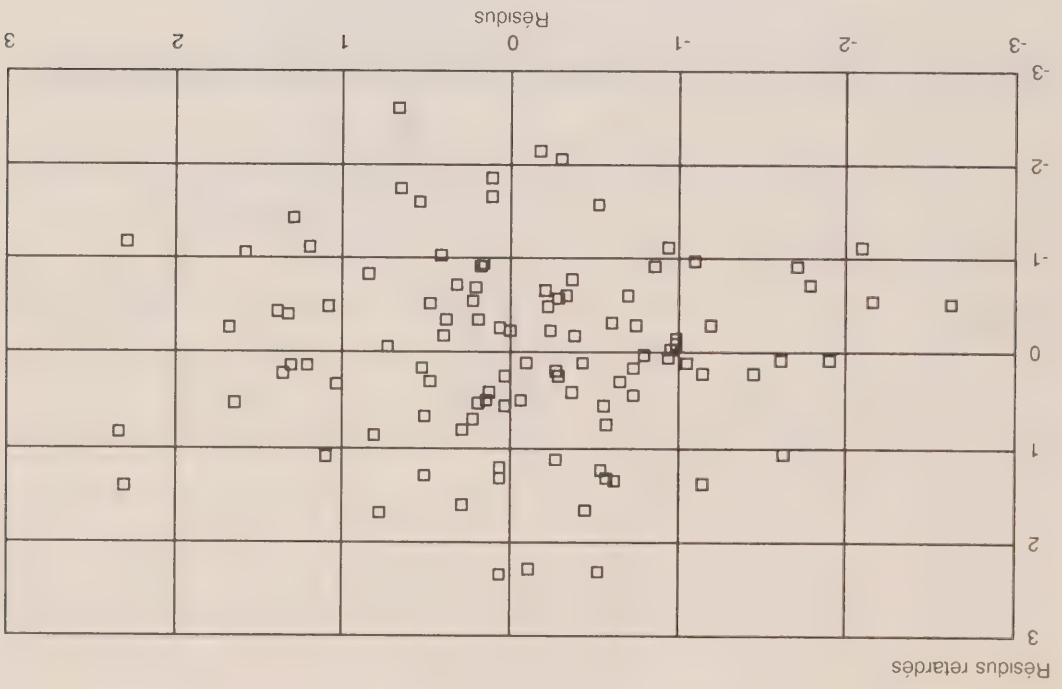
Nous cherchons à déterminer de quelle manière le nombre des naissances est lié au temps et à l'espace et, plus particulièrement, à découvrir des régimes de fécondité selon les périodes et les régions et, peut-être, des tendances inédites. L'étude comporte essentiellement deux volets. Nous présentons tout d'abord une analyse à pondération locale de données agrégées; ensuite,

¹ David R. Brillinger, Département de statistique, Université de Californie, Berkeley, CA, 94720, É.-U.

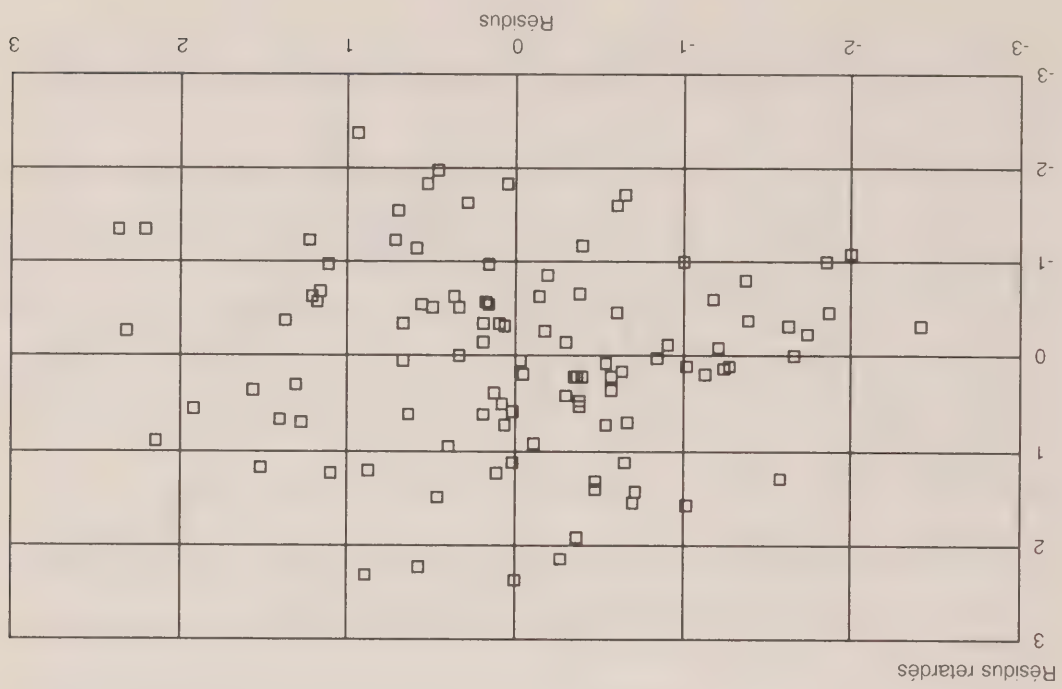
BIBLIOGRAPHIE

- ANSLEY, C.F., et KOHN, R. (1985). A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. *Journal of Statistical Computation and Simulation*, 21, 135-169.
- ANSLEY, C.F., et KOHN, R. (1986). Prediction mean squared error for state space models with estimated parameters. *Biometrika*, 73, 467-473.
- BELL, W.R., et HILLMER, S.C. (1987). Time Series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.
- BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées - Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.
- BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, séries B*, 35, 61-68.
- BROWN, R.L., DURBIN, J., et EVANS, J.M. (1975). Techniques for testing the consistency of regression relationships over time. *Journal of the Royal Statistical Society, séries B*, 37, 149-163.
- HARVEY, A.C., et DURBIN, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society, séries A*, 149, 187-222.
- HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- HAUSMAN, J.A., et WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.
- JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, séries B*, 42, 221-226.
- KOHN, R., et ANSLEY, C.F. (1986). Estimation, prediction and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, 81, 751-761.
- LEE, H. (1990). Estimation of panel correlations for the Canadian Labour Force Survey. *Survey Methodology*, 16, 297-306.
- MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. *Ph. D. Thesis*, Iowa State University, Ames, Iowa.
- RAO, J.N.K., SRINATH, K.P., et QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. Dans *Panel Surveys* (éds D. Kasprzyk, G. Duncan, G. Kalitin and M.P. Singh), New York: Wiley, 457-479.
- SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F., et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.
- TUNNICLIFFE-WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society, séries B*, 51, 15-27.

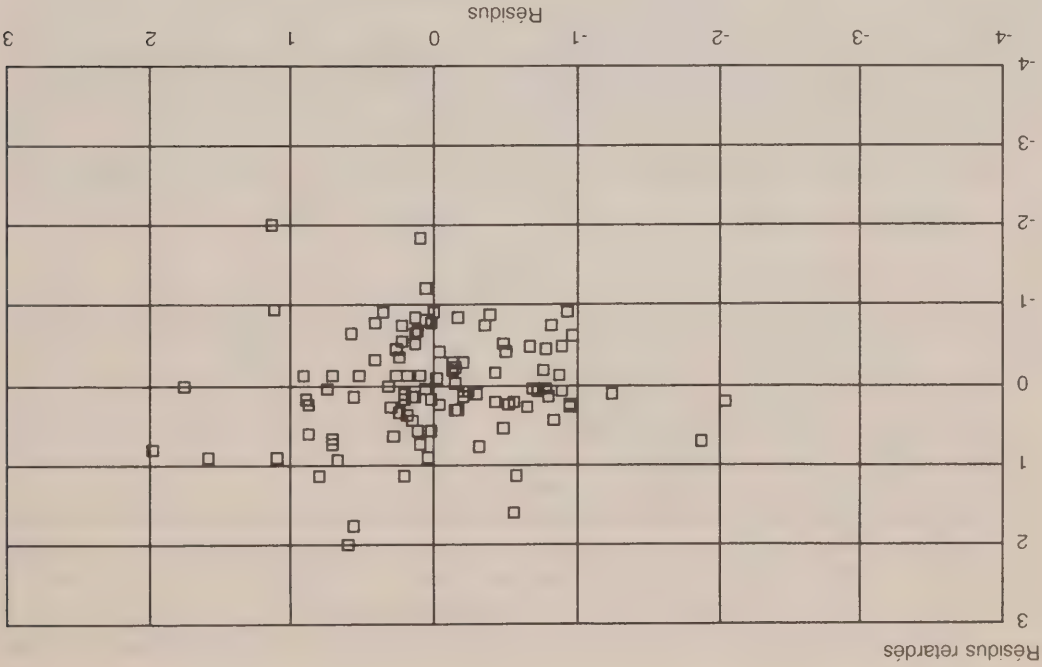




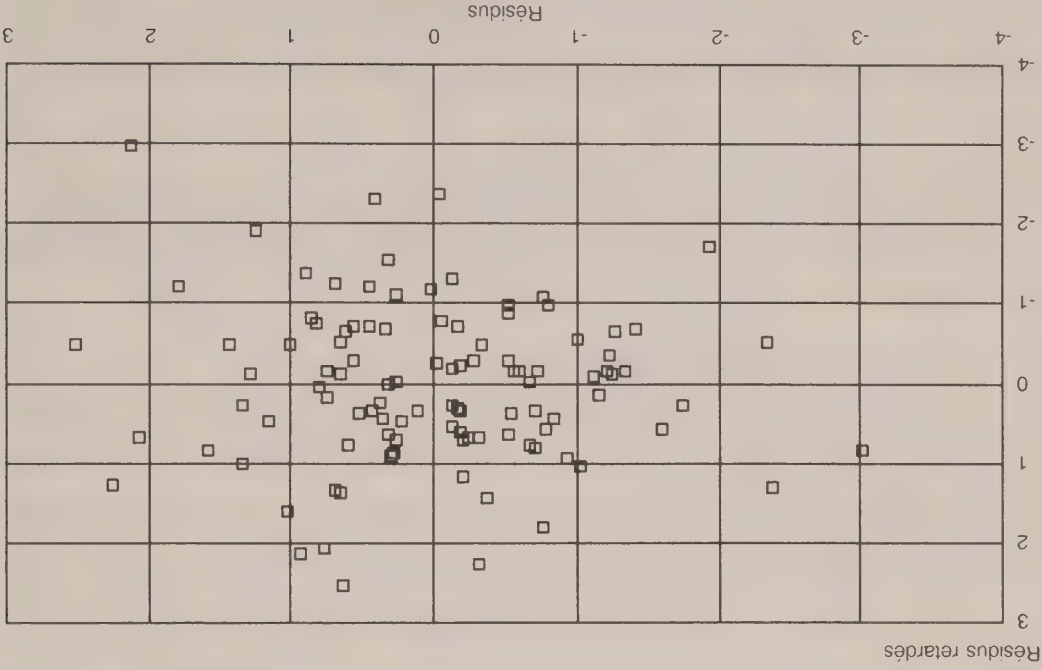
Graphique 3a Erreurs de prévision une étape à l'avance – Ile-du-Cap-Breton
(compte tenu de l'erreur de sondage)



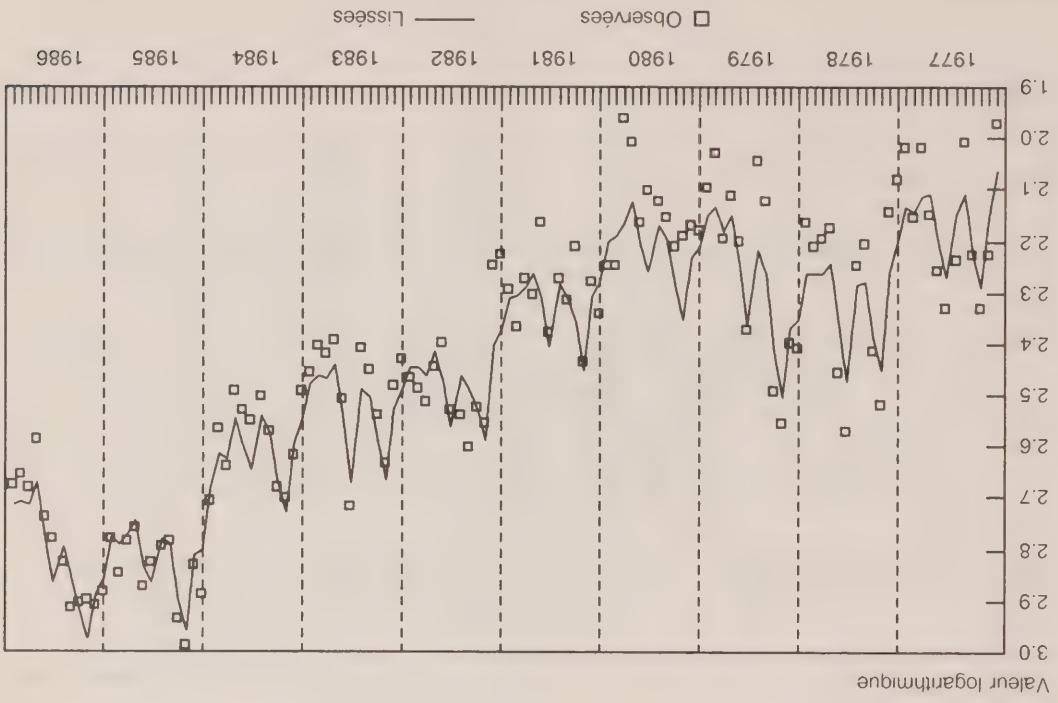
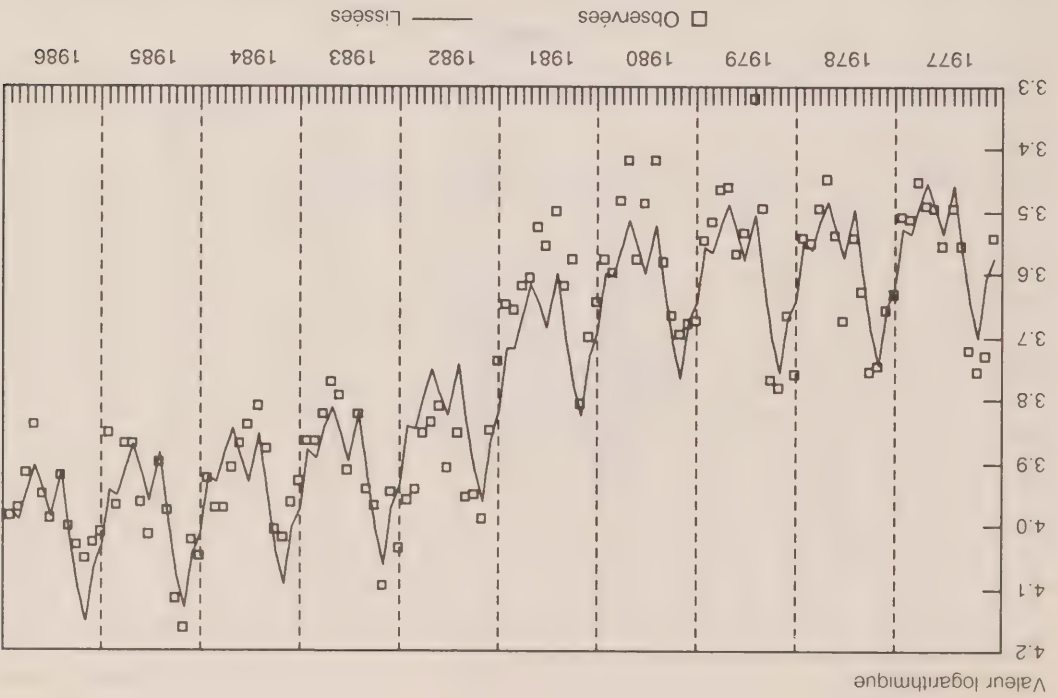
Graphique 3b Erreurs de prévision une étape à l'avance – Ile-du-Cap-Breton
(compte non tenu de l'erreur de sondage)



Graphique 2a Erreurs de prévision une étape à l'avance – Nouvelle-Ecosse
(compte tenu de l'erreur de sondage)



Graphique 2b Erreurs de prévision une étape à l'avance – Nouvelle-Ecosse
(compte non tenu de l'erreur de sondage)



Le tableau 1 renferme aussi les estimations pour la Nouvelle-Ecosse. Nous remarquons une réduction beaucoup plus forte de la variance estimée du modèle (lorsqu'on tient compte de l'erreur d'échantillonnage) dans le cas de la Nouvelle-Ecosse que dans celui de l'Île-du-Cap-Breton. Autre observation importante: la valeur estimée du paramètre autorégressif est très différente d'un modèle à l'autre. Les résultats montrent que ce paramètre a une grande importance dans chaque modèle. Selon le modèle qui ne tient pas compte de l'erreur d'échantillonnage, la valeur estimée du paramètre autorégressif est -0.296 tandis que selon l'autre modèle, elle est 0.862 . De toute évidence, ces deux valeurs appellent des interprétations tout à fait différentes. Le graphique 1a montre en superposition, pour la Nouvelle-Ecosse, les estimations lissées et les données originales pour le modèle qui tient compte de l'erreur d'échantillonnage. Le graphique 1b montre la même chose pour l'Île-du-Cap-Breton. Ce qui ressort le plus de ces graphiques est l'effet de la récession de 1981 sur les estimations lissées. Avant la récession, le modèle tend à surestimer le nombre de chômeurs alors qu'après, il tend à le sous-estimer.

5.2 Vérification du modèle

Nous avons produit des graphiques mettant en rapport les résidus récuratifs généralisés (définis dans la section 4.3) et les résidus récuratifs généralisés retardés pour tous les modèles. Les graphiques 2a et 2b concernent les deux modèles pour la Nouvelle-Ecosse. Notons que la dispersion autour de l'origine est moins grande dans le graphique 2a que dans le graphique 2b, ce qui dénote un meilleur ajustement lorsqu'on tient compte de l'erreur de sondage. Les graphiques 3a et 3b concernent les deux modèles pour l'Île-du-Cap-Breton. Leur similitude est frappante. Toutefois, aucun des quatre graphiques ne vient contredire l'hypothèse de normalité qui est à la base de chaque modèle.

Pour vérifier si les modèles n'ont pas subi de changement structurel, on peut faire la somme cumulée des résidus récuratifs et la représenter par un "graphique des sommes cumulées". Alors que les tests décrits dans Brown, Durbin et Evans (1975) n'ont pas été concluants, le graphique des sommes cumulées donne à penser qu'il s'est produit un changement structurel. Le graphique 4a, qui reproduit la somme cumulée des résidus pour la Nouvelle-Ecosse, montre très clairement que les résidus sont généralement négatifs avant la récession de 1981, ce qui signifie que les prédicteurs du modèle sont trop forts. Durant la récession, le modèle produit surtout des résidus positifs, ce qui dénote des prédicteurs trop faibles. Le graphique 4b reproduit la somme cumulée des résidus pour l'Île-du-Cap-Breton. Nous pouvons voir d'après ce graphique que le modèle qui tient compte de l'erreur de sondage subit un changement structurel plus tôt dans la période étudiée.

En conclusion, nous constatons que les modèles peuvent être améliorés. En ajoutant une variable de régression qui représenterait les changements structurels observés grâce au graphique des sommes cumulées, on pourrait pousser plus loin l'analyse dans le même cadre général. On étudierait actuellement la forme que pourrait prendre une telle variable.

5.3 Résumé

Ces exemples illustrent l'importance de tenir compte des erreurs de sondage dans certaines analyses chronologiques. À l'aide du filtre de Kalman modifié, nous avons pu élaborer pour un très grand nombre de modèles courants une méthode flexible pour l'estimation de paramètres, le lissage de données et la vérification de modèles.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à Bill Steele, de la Division des méthodes d'enquêtes sociales, pour avoir programmé l'algorithme décrit dans la section 4. Nous remercions également les responsables de l'Enquête sur la population active pour nous avoir fourni les séries de données. Enfin, l'arbitre et le rédacteur associé ont tous deux formulé de nombreux commentaires utiles qui sont venus rehausser cet article.

Le premier modèle que nous avons choisi pour les données de la Nouvelle-Ecosse, et qui fait abstraction de l'erreur d'échantillonnage, est un modèle ARMMI (1,1,0)(0,1,1)¹². Toutefois, comme le terme de moyenne mobile pour la composante saisonnière était estimé à un, nous avons utilisé un terme de régression déterministe pour tenir compte du mouvement saisonnier. Les 12 variables de régression contenant un terme linéaire et une variable auxiliaire pour chacun des 11 premiers mois. La variable auxiliaire pour un mois de référence prenait la valeur 1 pour ce mois, - 1 pour décembre et 0 pour les autres mois. Notons qu'il est superflu d'avoir une ordonnée à l'origine pour ce modèle puisque les différences premières des données sont ajustés.

Une analyse plus approfondie de ce modèle a révélé que la composante saisonnière de la moyenne mobile n'était pas nécessaire. Finalement, nous avons choisi, pour les données de la Nouvelle-Ecosse, un modèle ARMMI (1,1,0) avec un terme de régression déterministe. Nous avons utilisé le même modèle suivant l'hypothèse où il existe des erreurs d'échantillonnage. Nous nous sommes aussi servis du même modèle structurel pour la série de l'Ile-du-Cap-Breton. Le tableau 1 donne les valeurs estimées des paramètres. Les estimations qui font abstraction de l'erreur de sondage se trouvent dans les colonnes intitulées **Compte non tenu de l'erreur d'échantillonnage**. Si nous examinons tout d'abord les valeurs calculées pour l'Ile-du-Cap-Breton, nous voyons que les estimations par régression sont semblables d'un modèle à l'autre, ce qui était prévisible. Notons que les estimations d'erreur d'échantillonnage a une variance estimée beaucoup moins élevée. La colonne **Valeur T** renferme le quotient de la valeur estimée du paramètre par l'erreur type correspondante. Nous remarquons que la valeur *t* pour le paramètre autorégressif est très différente d'un modèle à l'autre (- 0.68 contre - 2.85). Cela implique que, pour les données de l'Ile-du-Cap-Breton, il faudrait choisir un modèle qui comporte uniquement un terme de régression déterministe lorsque l'erreur de sondage est prise en considération. En revanche, si l'erreur de sondage est exclue du modèle, on accordera trop d'importance au paramètre autorégressif.

Tableau 1

Valeurs estimées des paramètres - série du chômage 1977-1986

Paramètre	Nouvelle-Ecosse			Ile-du-Cap-Breton		
	Compte non tenu de l'erreur d'échantillonnage	Valeur T	Valeur estimée	Compte non tenu de l'erreur d'échantillonnage	Valeur T	Valeur estimée
Alpha	-0.296	- 3.23	0.862	-0.260	- 2.85	-0.231
Stigma	0.0597	-	0.0032	0.1049	-	0.0520
Tendance	0.00427	1.01	0.00420	0.00607	0.79	0.00598
Janvier	0.064	3.60	0.048	-0.007	-0.23	-0.003
Février	0.083	4.80	0.078	0.027	0.89	0.028
Mars	0.166	10.20	0.165	0.171	5.76	0.164
Avril	0.106	6.60	0.104	0.099	3.33	0.089
Mai	0.009	0.60	0.016	0.70	-0.008	-0.007
Juin	-0.101	- 6.00	-0.088	- 3.30	-0.029	-0.033
Juillet	-0.016	- 1.20	-0.014	-0.63	0.082	0.081
Août	-0.058	- 3.60	-0.062	- 2.37	-0.011	-0.009
Septembre	-0.106	- 6.60	-0.105	- 3.96	-0.104	-0.098
Octobre	-0.081	- 4.80	-0.071	- 3.08	-0.084	-0.069
Novembre	-0.026	- 1.80	-0.029	- 1.08	-0.063	-0.074

4.3 Résidus récurrents généralisés

Comme le soulignent Harvey et Durbin (1986), les résidus récurrents généralisés se prêtent bien à l'analyse diagnostique de modèles. Du point de vue du modèle d'espace d'états que nous avons défini, les résidus récurrents représentent l'écart entre la valeur observée et la valeur prévue une étape à l'avance, déterminée au moyen du filtre de Kalman. Ils peuvent être utilisés pour toutes les périodes t où $V_1(t + 1 | t) = 0$. Selon le modèle, ces résidus sont approximativement indépendants et distribués normalement. On peut leur appliquer une distribution normale de sorte que leur variance estimée soit égale à 1 selon le modèle. On peut alors faire des analyses diagnostiques semblables à celles que l'on trouve dans les modèles de régression classiques.

5. ANALYSE DES DONNÉES DE LA POPULATION ACTIVE

5.1 Estimation des paramètres

Pour illustrer ce dont nous venons de parler, nous allons nous servir de données de l'Enquête sur la population active du Canada (EPA). L'EPA est une enquête mensuelle avec groupes de renouvellement; chaque groupe (ou panel) correspond à un sixième de l'échantillon de ménages. Un panel demeure dans l'échantillon pendant six mois consécutifs tandis que les unités primaires d'échantillonnage sont renouvelées au bout de deux ans environ. L'échantillonnage se fait selon un plan stratifié à plusieurs degrés.

Les données en question sont le nombre mensuel de personnes en chômage pour la période de janvier 1977 à décembre 1986 dans la province de Nouvelle-Ecosse et la région de l'Île-du-Cap-Breton. Nous avons choisi la Nouvelle-Ecosse parce que les erreurs d'échantillonnage y sont moins élevées que dans les provinces plus grandes. Par ailleurs, nous avons choisi la région de l'Île-du-Cap-Breton parce qu'à cause de l'échantillon moindre, nous obtenons des estimations qui ont une variance relative plus élevée. Les graphiques 1a et 1b donnent le logarithme des séries pour la Nouvelle-Ecosse et l'Île-du-Cap-Breton respectivement. Les valeurs logarithmiques ont servi de données d'entrée.

Lee (1990) a estimé les autocorrélations de l'erreur de sondage pour la Nouvelle-Ecosse jusqu'à un décalage de onze. Nous avons calculé les coefficients du processus ARMA (m, n) pour l'erreur de sondage, défini en (2.7), en apartiant ces autocorrélations. Nous avons pu obtenir un bon ajustement à l'aide d'un modèle ARMA (3,6). Voici les coefficients en question:

$\phi_1 =$	0.2575	$\psi_1 =$	-0.1847
$\phi_2 =$	-0.3580	$\psi_2 =$	-0.5873
$\phi_3 =$	-0.6041	$\psi_3 =$	0.3496
$\psi_4 =$	0.0647		
$r^2 =$	0.7246	$\psi_5 =$	0.0982
$\psi_6 =$	0.0347		

Le facteur k_i dans (2.6) représente l'erreur type estimée de l'estimation et est calculé en appliquant la formule d'approximation de Taylor aux logarithmes. Nous avons tout d'abord ajusté une série de modèles aux données de la Nouvelle-Ecosse en supposant qu'il n'y avait pas d'erreur d'échantillonnage. Nous avons ensuite repris l'opération en incorporant cette fois le modèle prévoyant des erreurs de sondage. Dans cette deuxième phase, nous avons pu aussi calculer des valeurs lissées pour les estimations d'enquête et comparer les erreurs types de ces estimations lissées à celles de la série originale.

La fonction de vraisemblance a été maximisée à l'aide d'une version modifiée de la méthode de la fonction de caractérisation. Cette version modifiée prévoyait des pas variables. À chaque itération, on calculait la valeur de la fonction de vraisemblance pour le pas précédent et pour le pas courant multiplié et divisé par une constante préalable (en l'occurrence 1.1). Dans un deuxième temps, on choisissait parmi les trois valeurs celle qui maximisait la fonction de vraisemblance. À chaque fois, on vérifiait si les paramètres se situaient dans les limites voulues. Cette condition était réalisée si la matrice des covariances initiales du vecteur d'états était semi-définie positive. Si tel n'était pas le cas, on divisait de nouveau le pas par la constante et on répétait l'opération.

Pour estimer la matrice des variances des paramètres estimés, on s'est servi de l'inverse de la matrice d'information de Fisher. Ce calcul se fait facilement puisqu'on connaît les dérivées premières de la fonction de vraisemblance.

4.2 Estimation des valeurs lissées

On obtient des estimations lissées comme celles définies en (3.4) en rendant nulle la composante du vecteur d'états qui représente l'erreur de sondage. Cependant, cette simplification ne nous dit pas comment estimer la variance d'estimations lissées. Lorsqu'on veut calculer l'erreur type d'une estimation lissée, il faut tenir compte de ce que les paramètres inconnus ont été estimés à l'aide des données observées (surtout lorsqu'il s'agit d'une série de données relativement courte); voir Jones (1979).

Pour connaître la variance de $g'z_t$, il suffit de calculer la variance de $z_T - \hat{m}(T|T)$, où $\hat{m}(T|T)$ est la valeur estimée de $m(T|T)$ aux valeurs estimées des paramètres, car le vecteur d'états comprend désormais $g'z_t$. Or,

$$z_T - \hat{m}(T|T) = [z_T - m(T|T)]$$

$$+ [m(T|T) - \hat{m}(T|T)]. \quad (4.2)$$

Le premier terme du membre de droite de l'équation (4.2) a pour variance conditionnelle $V(T|T) = V_0(T|T)$, en supposant que $V_1(T|T) = 0$. Le second terme du membre de droite est un terme d'erreur et est indépendant du premier terme puisqu'il ne dépend que des observations y . Si nous soumettons le second terme du membre de droite à un développement de Taylor par rapport aux valeurs réelles des paramètres et que nous ne tenons pas compte des termes de degré supérieur, nous obtenons

$$m(T|T) - \hat{m}(T|T) = \left[-\frac{\partial \hat{m}(T|T)}{\partial \phi} \right] (\hat{\phi} - \phi), \quad (4.3)$$

où ϕ est le vecteur des paramètres inconnus et $\hat{\phi}$, l'estimation correspondante. Par conséquent, la variance asymptotique de (4.2) est approximativement

$$\text{Var}[z_T - \hat{m}(T|T)] = V_0(T|T)$$

$$+ \left[\frac{\partial \hat{m}(T|T)}{\partial \phi} \right] V_\phi \left[\frac{\partial \hat{m}(T|T)}{\partial \phi} \right], \quad (4.4)$$

où V_ϕ est la matrice des covariances pour les paramètres inconnus. On estime la variance définie en (4.4) au moyen des valeurs estimées des paramètres. Cette méthode est identique à celle présentée par Ansley et Kohn (1986).

Or, on suppose que u_{-} est un processus ARMA stationnaire, de sorte qu'il est possible de déterminer la matrice des covariances correspondante à l'aide de l'équation (2.5). Par ailleurs, on suppose que w_{-} est distribué selon $N(0, \kappa I)$ et est indépendant de u_{-} . Puisque (u_{-}', w_{-}') est une combinaison linéaire non singulière de θ_{-} , il est possible de calculer la matrice des covariances pour θ_{-} . En nous servant de la forme de l'expression (3.5) pour z_0 , nous pouvons calculer la matrice des covariances initiales. Il convient de souligner que lorsque d et D sont nuls, de sorte qu'il ne se fait pas de calculs de différence dans le modèle, w_{-} devient un vecteur nul et nous avons $u_{-} = \theta_{-}$.

3.3 Modèle pour les observations

Dans la section 2, nous avons supposé que $e_t = k_t' \omega_t$, où ω_t est un modèle ARMA(m, n). Par conséquent, il est clair, d'après l'analyse de la section 3.2, que l'on peut représenter e_t sous forme de modèle d'états où $h_t' = (k_t', 0, \dots, 0)'$, et $e_t = h_t' z_t$. On peut représenter la composante de régression sous la même forme en ajoutant γ au vecteur d'états et en supposant au départ que γ a une moyenne nulle et une covariance κI . Notons que γ est constant dans l'équation de transition. Puisque nous pouvons représenter chaque composante de l'équation (2.1) par un modèle d'espace d'états, il est facile de combiner les différents modèles en un seul en formant un vecteur unique avec les vecteurs d'états définis pour chaque composante. L'équation d'observation est alors la somme des trois composantes.

4. ESTIMATION DU MODELE D'ESPACE D'ETATS

4.1 Estimation des paramètres

Les paramètres inconnus de ce modèle sont σ^2 et les coefficients de $\lambda(B)$, $\alpha(B)$, $v(B)$ et $\beta(B)$. Dans l'opération de maximisation numérique décrite ci-dessous, nous avons transformé σ^2 en $\log(\sigma^2)$ pour que les valeurs négatives des paramètres ne posent pas de difficultés. Le modèle pour le vecteur d'observations $y = (y_1, y_2, \dots, y_T)'$ défini dans la section 3 est équivalent à

$$y = M\eta + \zeta, \tag{4.1}$$

où η est un vecteur à j dimensions distribué selon $N(0, \kappa I)$, ζ est un vecteur à T dimensions distribué selon $N(0, W)$ et M est une matrice fixe $T \times j$. Notons que η renferme des constantes inconnues, y compris les coefficients de régression; W est une fonction des paramètres du modèle ARMA; M est une fonction de la structure des calculs de différence. Kohn et Ansley (1986) ont recommandé de maximiser la limite du produit de $\kappa^{j/2}$ par la fonction de vraisemblance pour les données, lorsque κ tend vers l'infini. Il est possible de montrer que cette limite de la fonction de vraisemblance équivaut à la fonction de vraisemblance marginale de $y - M\eta$, où η est l'estimation la plus vraisemblable de η lorsque M et W sont connues. Tunncliffe-Wilson (1989) a montré que le jacobien de la transformation des données y vers $(\eta, y - M\eta)$ ne dépend pas des paramètres de modèle de W lorsque M est connue. Ansley et Kohn (1985) ont montré que M ne dépend pas des paramètres inconnus. Le calcul de la fonction de vraisemblance marginale est plus simple avec le filtre de Kalman modifié qu'avec la méthode proposée par Tunncliffe-Wilson. La méthode que nous avons utilisée pour cette analyse permet de calculer la fonction de vraisemblance marginale ainsi que les dérivées premières de cette fonction par rapport aux paramètres inconnus. Cela implique le calcul des dérivées premières des conditions initiales et de $m(t | t')$ et des composantes de $V(t | t')$ pour $t = t'$ et $t = t' + 1$. Tous les calculs ont été faits à l'aide de PROC IML du SAS.

Ensuite, on ajoute au vecteur d'états z_t l'élément $z_{t,r+1} = g'_t z_t$, et $m(t | t)$ et $V(t | t)$ sont rajustées en conséquence. Enfin, on modifie la matrice F dans (3.1b) de manière à y inclure l'équation $z_{t+1,r+1} = z_{t,r+1}$. Une fois ces modifications faites, on peut utiliser de nouveau le filtre de Kalman modifié de manière que le dernier élément de $m(T | T)$ représente l'espérance conditionnelle de $g'_T z_T$ étant donné toutes les observations, y_1, y_2, \dots, y_T . De même, le dernier élément diagonal de $V(t | t)$ représente la variance conditionnelle de $g'_t z_t$. On peut généraliser cette méthode en y introduisant un nombre indéterminé d'estimations lissées ainsi que les covariances conditionnelles correspondantes. Dans des applications, on pourrait ne pas être en mesure de calculer les estimations lissées pour un grand nombre de périodes à cause d'une capacité de stockage limitée.

3.2 Modèle pour θ

Harvey et Phillips (1979) ont décrit une méthode permettant d'exprimer le modèle ARMMI (2.4) sous forme de modèle d'états (équ. 3.1). La dimension de z_t est $r = \max(p + d + sP + sD, q + sQ)$. En ajoutant des zéros à $A = (A_1, \dots, A_{p+d+sP+sD})$ ou à $b = (b_1, \dots, b_{q+sQ})$ de manière à obtenir des vecteurs de dimension r , on peut exprimer le modèle ARMMI sous la forme donnée en (3.1), où $h'_t = (1, 0, \dots, 0)$, $G'_t = (1, -b_1, \dots, -b_{r-1})$ et

$$F = \begin{bmatrix} A_1 & \vdots & A_{r-1} & \left| \begin{array}{c} A_r \\ 0' \end{array} \right. \\ \hline I_{r-1} & & & \end{bmatrix},$$

où I_{r-1} est la matrice unité $(r-1) \times (r-1)$ et $0'$ est un vecteur ligne formé de zéros.

Dans cette formulation, le vecteur d'états $z_t = (z_{1t}, \dots, z_{rt})'$ est défini

$$z_{it} = A_i \theta_{t-i-1} + A_{i+1} \theta_{t-i-2} + \dots + A_r \theta_{t-(r-i+1)} - b_{i-1} \epsilon_{t-i-1} - b_i \epsilon_{t-i} - \dots - b_{r-1} \epsilon_{t-(r-i)} \quad (3.5)$$

pour $i = 2, 3, \dots, r$ et $z_{1t} = \theta_t$.

Pour compléter la spécification du modèle pour $\{\theta_t\}$, il reste à définir les conditions initiales pour z_0 . Celles-ci sont données dans Ansley et Kohn (1985) et nous en faisons un résumé ci-dessous.

D'après l'équation (2.5), $\{u_t\}$ est un processus ARMA. Nous posons

$$\theta_- = (\theta_0, \theta_{-1}, \dots, \theta_{-S})',$$

où $S = \max(0, p + sP + d + sD - 1)$. De plus,

$$u_- = (u_0, u_{-1}, \dots, u_{-R})',$$

où $R = \max(0, p + sP - 1)$. Enfin,

$$w_- = (\theta_{-R-1}, \theta_{-R-2}, \dots, \theta_{-S})',$$

lorsque $S > R$.

3. FORMULATION D'UN MODELE D'ESPACES D'ETATS

3.1 Formulation générale

Le modèle décrit dans la section 2 peut être redéfini comme un modèle d'espace d'états avec des conditions initiales partiellement diffusées. Cela présente un certain nombre d'avantages. Par exemple, le modèle d'espace d'états permet de calculer, au moyen d'un filtre de Kalman modifié, une fonction de vraisemblance marginale, que l'on peut maximiser dans le but d'estimer des paramètres inconnus. Il facilite aussi le lissage des estimations originales en éliminant de ces estimations l'erreur de sondage.

Dans le modèle d'espace d'états, deux processus se déroulent simultanément. Le premier, qui est le système d'observation, décrit en détail comment les observations dépendent de l'état des paramètres du processus dans la période observée. Le second, qui est le système de transition, décrit en détail l'évolution des paramètres.

Pour les modèles d'espace d'états considérés dans cet article, l'équation d'observation s'écrit:

(3.1a) $y_t = h_t'z_t$

et l'équation de transition est:

(3.1b) $z_t = Fz_{t-1} + G\xi_t$,

où z_t est un vecteur d'états ($r \times 1$) et h_t , un vecteur fixe ($r \times 1$). Dans l'équation de transition, F est une matrice fixe ($r \times r$), G est une matrice fixe ($r \times m$) et les ξ_t sont des vecteurs normaux indépendants de moyenne nulle et de covariance U .

La dernière étape de la spécification du modèle d'espace d'états consiste à définir les conditions initiales pour z_0 . Dans cet article, nous allons reprendre les conditions posées dans Kohn et Ansley (1986). De façon générale, nous supposons que z_0 suit une distribution normale partiellement diffuse à r dimensions avec comme moyenne $m(0|0) = 0$ et comme matrice des covariances $V(0|0)$, où

(3.2) $V(0|0) = \kappa V_1(0|0) + V_0(0|0)$

pour des valeurs κ élevées. La matrice $V_1(0|0)$ définit la portion diffuse de la distribution a priori. Dans la section 3.2, nous expliquons comment obtenir $V_1(0|0)$ et $V_0(0|0)$ pour notre modèle.

Nous désignons par $m(t|t')$ la moyenne conditionnelle de z_t étant donné les observations allant jusqu'à la période t' incluse, et par $V(t|t')$, sa variance conditionnelle, où

(3.3) $V(t|t') = \kappa V_1(t|t') + V_0(t|t')$.

Les formules récursives pour les cas où $t = t'$ et $t = t' + 1$ sont définies dans Kohn et Ansley (1986). Ceux-ci les désignent comme le filtre de Kalman modifié.

Comme le modèle défini en (2.1) contient une composante d'erreur $\{e_t\}$, il est intéressant de voir ce que donnerait une estimation du modèle sans l'erreur de sondage, c'est-à-dire:

(3.4) $y_t(\text{lissée}) = x_t'\gamma + \theta_t$.

Lorsqu'on peut exprimer le membre de droite de l'équation (3.4) par $g_t'z_t$, pour une valeur g_t' quelconque, on peut calculer la moyenne et la variance conditionnelles de la combinaison linéaire $g_t'z_t$, étant donné toutes les observations, au moyen du filtre de Kalman modifié. À cette fin, on applique les récursions jusqu'à la période t pour déterminer $m(t|t)$ et $V(t|t)$.

(2.3)
$$a(B)\Delta(B)\theta_i = b(B)\epsilon_i,$$

(2.4)
$$A(B)\theta_i = b(B)\epsilon_i,$$

(2.5)
$$a(B)u_i = b(B)\epsilon_i.$$

Considérons maintenant les erreurs de sondage $\{\epsilon_i\}$ de l'équation (2.1). Nous supposons que les échantillons de l'enquête à passages répétés sont suffisamment grands pour que les erreurs de sondage puissent être représentées approximativement par une distribution normale multidimensionnelle. Dans la plus simple des hypothèses, c'est-à-dire lorsqu'il n'y a pas de participation répétée des unités et que les fractions de sondage sont faibles, on peut supposer que les ϵ_i sont indépendantes. Dans les enquêtes par panel, les erreurs de sondage sont le plus souvent corrélées. Comme, dans ce genre d'enquêtes, les corrélations d'un passage à l'autre deviennent nulles après la suppression de panels, un simple processus de moyennes mobiles peut servir à décrire les erreurs de sondage. Par ailleurs, si c'est un échantillon aléatoire d'unités qui est renouvelé à chaque passage de l'enquête, un processus autorégressif pur est probablement ce qui décrit le mieux les erreurs de sondage. Il existe aussi des modèles plus complexes. Par exemple, dans un plan d'échantillonnage à deux degrés, une partie des unités du premier degré peuvent être remplacées aléatoirement à chaque passage tandis que les unités du second degré peuvent constituer un échantillon avec renouvellement. Dans un tel cas, le processus autorégressif à moyennes mobiles (ARMA) pourrait être le meilleur moyen de décrire les erreurs de sondage, comme le laissent croire Scott, Smith et Jones (1977).

Dans cet article, nous supposons que les erreurs de sondage sont décrites par l'expression

(2.6)
$$\epsilon_i = k_i \omega_i,$$

où $\{\omega_i\}$ est un processus ARMA (m,n) défini par l'équation

(2.7)
$$\phi(B)\omega_i = \psi(B)\eta_i$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_m B^m,$$

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_n B^n.$$

Les η_i sont indépendants et identiquement distribués selon $N(0,\tau^2)$. Le facteur k_i a été inclus dans l'équation (2.6) pour tenir compte des variances non homogènes lorsque la fonction d'autocorrélation est homogène dans le temps.

Dans le modèle que nous venons de décrire, nous supposons que τ^2 , k_i et les coefficients de $\phi(B)$ et de $\psi(B)$ peuvent être estimés directement à partir des données d'enquête à l'aide de méthodes fondées sur un plan. En revanche, les autres paramètres sont réputés inconnus. Ceux-ci comprennent γ , σ^2 , et les coefficients de $\lambda(B)$, $\alpha(B)$, $\nu(B)$ et $\beta(B)$. Les x_i du terme de régression sont supposés connus.

on peut lisser les estimations d'enquête à l'aide de méthodes empiriques de Bayes de manière qu'elles intègrent les caractéristiques du modèle. Nous définissons des intervalles de confiance pour ces valeurs lissées en nous servant de la méthode décrite par Ansley et Kohn (1986). Bell et Hillmer (1987) ont employé un modèle semblable mais avec des conditions initiales plus rigides, c'est-à-dire des conditions qui rendent ce modèle moins propre à tenir compte des termes de régression ou des valeurs manquantes (tout en conservant l'approche de la fonction de vraisemblance marginale). Nous donnons un exemple de ce modèle dans la section 5 en nous servant de données sur le chômage tirées de l'Enquête sur la population active du Canada. Cet exemple montre en quoi le fait de tenir compte des erreurs de sondage influe sur les estimations des paramètres du modèle. Nous calculons une estimation lissée du processus correspondant selon les hypothèses du modèle. Nous calculons aussi des résidus récuratifs et recourons à des techniques de vérification pour évaluer les divers modèles.

2. LE MODÈLE

Supposons que nous avons une série d'estimations ponctuelles d'une caractéristique de la population, y_1, y_2, \dots, y_T , qui proviennent d'une enquête à passages répétés. Nous supposons que y_t peut être décomposé en trois éléments

$$(2.1) \quad y_t = x_t' \gamma + \theta_t + e_t,$$

où $x_t' \gamma$ est un terme de régression déterministe, θ_t est un paramètre de population qui répond à un modèle de série chronologique et e_t est l'erreur de sondage, dont l'espérance est nulle par hypothèse. Nous allons tout d'abord décrire un modèle autorégressif à moyennes mobiles intégré pour $\{\theta_t\}$. Posons B comme l'opérateur de décalage (ou de retard), $\nabla = 1 - B$ et $\nabla_s = 1 - B^s$, où s est la période saisonnière. Nous définissons les fonctions polynomiales suivantes:

$$\begin{aligned} \lambda(B) &= 1 - \lambda_1 B - \lambda_2 B^2 - \dots - \lambda_p B^p, \\ \alpha(B) &= 1 - \alpha_1 B - \alpha_2 B^2 - \dots - \alpha_p B^p, \\ v(B) &= 1 - v_1 B - v_2 B^2 - \dots - v_q B^q, \\ \beta(B) &= 1 - \beta_1 B - \beta_2 B^2 - \dots - \beta_q B^q. \end{aligned}$$

et

Le modèle ARMMI (p, d, q) (P, D, Q)_s pour $\{\theta_t\}$ est défini

$$(2.2) \quad \lambda(B^s) \alpha(B) \nabla^d \nabla_s^d \theta_t = v(B^s) \beta(B) e_t,$$

où les e_t sont indépendants et identiquement distribués selon $N(0, \sigma^2)$. Nous définissons $a(B) = \lambda(B^s) \alpha(B)$ comme un polynôme de degré ($p + sP$); $\Delta(B) = \nabla^d \nabla_s^d$ comme un polynôme de degré ($d + sD$); $b(B) = v(B^s) \beta(B)$ comme un polynôme de degré ($q + sQ$); $A(B) = a(B) \Delta(B)$ comme un polynôme de degré ($p + d + sP + sD$); $u_t = \Delta(B) \theta_t$, comme un processus ARMA ($p + sP, q + sQ$). En conséquence, nous pouvons représenter l'équation (2.2) de diverses façons:

Méthode pour l'analyse des modèles ARMMI

DAVID A. BINDER et J. PETER DICK¹

RÉSUMÉ

Le modèle ARMMI est souvent utilisé pour l'analyse des modèles de séries chronologiques. Toutefois, ce genre d'analyse fait souvent abstraction des erreurs contenues dans les données d'enquête. Par l'intermédiaire de modèles d'états comportant des conditions initiales partiellement diffusées, les auteurs montrent comment estimer les paramètres inconnus du modèle ARMMI à l'aide des méthodes du maximum de vraisemblance. En outre, ils montrent qu'il est possible de lisser les estimations d'enquête à l'aide d'un modèle empirique de Bayes et de faire une vérification du modèle ARMMI. Enfin, ils appliquent ces techniques à une série sur le chômage tirée de l'Enquête sur la population active.

MOTS CLÉS: Filtre de Kalman; fonction de vraisemblance partielle; lissage de données.

1. INTRODUCTION

Les méthodes d'analyse chronologique sont largement utilisées aujourd'hui pour analyser les données des enquêtes à passages répétés. La plupart de ces méthodes supposent soit l'absence d'erreurs de sondage ou bien l'indépendance de ces erreurs, si elles existent. Or, dans le cas d'enquêtes à passages répétés, où certaines unités d'échantillonnage reviennent d'une fois à l'autre, les erreurs de sondage peuvent être corrélées dans le temps.

Un modèle fréquemment utilisé pour l'analyse chronologique est le modèle autorégressif à moyennes mobiles intégré (ARMMI), qui fait l'objet de cet article. Nous allons voir comment introduire dans l'analyse les erreurs de sondage (qui sont peut-être corrélées). En particulier, nous allons envisager le cas où l'on peut supposer que les erreurs de sondage suivent un processus autorégressif à moyennes mobiles (ARMA) jusqu'à une constante multiplicative.

Lorsqu'on suppose un modèle de ce genre pour décrire le mouvement des caractéristiques de la population, il est possible de déterminer l'estimateur à erreur quadratique moyenne minimum (ou estimateur linéaire de Bayes) d'une caractéristique à un moment précis. Cet estimateur admet la structure de modèle que les estimateurs classiques, comme l'estimateur linéaire non biaisé à variance minimum, n'admettent pas. Lorsque les paramètres du modèle sont estimés à l'aide des données d'enquête, il s'agit alors d'estimateurs empiriques de Bayes.

Blight et Scott (1973), Scott et Smith (1974), Scott, Smith et Jones (1977), Jones (1980), Rao, Srinath et Quenneville (1989) et d'autres encore ont examiné les conséquences de certains modèles stochastiques pour les moyennes de population dans le temps. Hausman et Watson (1985), pour leur part, introduisent un modèle d'erreurs d'observation dans le processus de désaisonnalisation courant. Mizazaki (1985) suppose que l'on peut modéliser les erreurs de sondage au moyen d'un simple processus de moyennes mobiles. Binder et Dick (1989) généralisent ces résultats à l'aide de modèles d'espace d'états et de filtres de Kalman. Dans cet article, nous élargissons le cadre de l'analyse en considérant le cas où le calcul de différences pour la série originale des moyennes de population se traduit par un modèle ARMA. À cette fin, nous utilisons une version modifiée de la méthode du filtre de Kalman, proposée par Kohn et Ansley (1986). Pour estimer les paramètres inconnus, nous maximisons la fonction de vraisemblance marginale par la méthode de caractérisation. La méthode de Kohn et Ansley renferme aussi des mécanismes pour les données manquantes. Nous allons voir aussi comment

¹ D.A. Binder, Division des méthodes d'enquêtes-entreprises et J.P. Dick, Division des méthodes d'enquêtes sociales, Statistique Canada, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 163-175.
- PFEFFERMANN, D., et BARNARD, C. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* (à paraître).
- PFEFFERMANN, D., BURCK, L., et BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *L'analyse des données dans le temps*. Ottawa: Statistique Canada (à paraître).
- PFEFFERMANN, D., et SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *Revue Internationale de Statistique*, 53, 37-59.
- ROSENBERG, B. (1973a). The analysis of cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement*, 2, 399-428.
- ROSENBERG, B. (1973b). A survey of stochastic parameter regression. *Annals of Economic and Social Measurement*, 2, 381-397.
- SÄRNDA, C.E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHWEPPE, F. (1965). Evaluation of likelihood functions for gaussian signals. *IEEE Transactions on Information Theory*, 11, 61-70.
- SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (éd.) N.K. Nawboodivi, New York: Academic Press, 201-216.
- SWAMY, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.
- TILLER, R. (1989). A Kalman filter approach to labour force estimation using survey data. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association* (à paraître).
- WATSON, M.W., et ENGLE, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23, 385-400.

D'après la définition de la matrice $F_{(P)}'$ (voir au-dessous de l'équation 4.3), nous pouvons réécrire l'équation (A6) de la façon suivante:

$$P_{(A)}' = \tilde{Q} P_{(A)}'^{l-1} - P_{(A)}'^{l-1} Z_{(A)}' K_{(P)}' + K_{(P)}' F_{(P)}' K_{(P)}' \quad (A7)$$

$$+ K_{(P)}' (\Sigma_{(A)}' - \Sigma_{(P)}') K_{(P)}'$$

par des transformations algébriques simples, l'expression ci-dessus devient l'équation (4.4).

BIBLIOGRAPHIE

ANDERSON, B.O.D., et MOORE, J.B. (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

ANSLEY, C.F., et KOHN, R. (1986). Prediction mean squared error for State Space models with estimated parameters. *Biometrika*, 73, 467-473.

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées - Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, Vol. 6, (éds.), P.R. Krishnaiah et C.R. Rao, Amsterdam: Elsevier Science, 187-211.

CHOUDHRY, G.H., et RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time* (à paraître).

COOLEY, T.F., et PRESCOTT, E.C. (1976). Estimation in the presence of stochastic parameter variation. *Econometrica*, 44, 167-184.

DIELMAN, T.E. (1983). Pooled cross-sectional and time series data: A survey of current statistical methodology. *The American Statistician*, 37, 111-122.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a State-Space model. *Journal of Econometrics*, 33, 388-397.

HARVEY, A.C. (1981). *Time Series Models*. Deddington, Oxford: Philip Allan.

HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

KITAGAWA, G., et GERSCH, W. (1984). A smoothness priors State-Space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79, 378-389.

JOHNSON, L.W. (1977). Stochastic parameter regressions: An annotated bibliography. *Revue Internationale de Statistique*, 45, 257-272.

JOHNSON, L.W. (1980). Stochastic parameter regression: An additional annotated bibliography. *Revue Internationale de Statistique*, 48, 95-102.

LA MOTTE, L.R., et McWHORTER, A. (1977). Estimation, testing and forecasting with random coefficient regression models. *Proceedings of the Section on Business and Economic Statistics, American Statistical Association*, 814-817.

MADDALA, G.S. (1977). *Econometrics*. Kogakusta: McGraw-Hill.

MEINHOLD, R.J., et SINGPURWALLA, N.D. (1983). Understanding the Kalman filter. *The American Statistician*, 37, 123-127.

ANNEXE

a) Détermination de l'équation (2.12)

Lorsque $\tilde{x}_{lki} = \tilde{x}_{lk}$, $\tilde{\Theta}_{lk} = \tilde{x}'_{lk} \tilde{\Theta}_{lk} = \tilde{x}'_{lk} \tilde{\Theta}_{lk}$ de sorte que $\tilde{\Theta}_l = (\tilde{\Theta}_{l1}, \dots, \tilde{\Theta}_{lk})' = Z'_l \tilde{\Theta}_l$. En outre, pour ce qui a trait au modèle de trajet aléatoire, la matrice T est la matrice unité et, d'après l'équation (3.1),

$$Z_l \tilde{\Theta}_l = Z_l P_{l|l-1} Z'_l + \Sigma_l F_{l-1}^{-1} (I - \Sigma_l F_{l-1}^{-1}) \tilde{Y}_l + \Sigma_l F_{l-1}^{-1} Z_l \tilde{\Theta}_{l-1} \quad (A1)$$

puisque $F_l = (Z_l P_{l|l-1} Z'_l + \Sigma_l)$. Supposons, pour des raisons de commodité, que $k = 1$ et posons

$$F_l = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \quad \text{et} \quad H_l = F_l^{-1} = \begin{bmatrix} h_{11} & \tilde{h}_l \\ h_{21} & \tilde{h}_l' \end{bmatrix} \quad \text{où} \quad f_{11} \quad \text{et} \quad h_{11}$$

sont des grandeurs scalaires, f_{12} et \tilde{h}_l' sont de dimension $[1 \times (K-1)]$ et F_{22} et H_{22} sont de dimension $[(K-1) \times (K-1)]$. En nous servant de la notation ci-dessus, nous pouvons déduire de l'expression (A1) l'équation suivante:

$$\Theta_{1l} = \left(1 - \frac{\sigma_1^2}{\sigma_2^2} h_{11} \right) Y_{1l} + \frac{\sigma_1^2}{\sigma_2^2} h_{11} (x'_{1l} \hat{\Theta}_{l-1,1}) - \frac{\sigma_1^2}{\sigma_2^2} \sum_{k=2}^K n_{1k} h_{11} \frac{h_{1k}}{h_{1k}} e_{1k} \quad (A2)$$

Désignons par $\tilde{f}_l' = (f_{12}, \dots, f_{1K}) = \tilde{f}_l' F_{22}^{-1}$ les coefficients de régression partielle dans la régression de e_{1l} par rapport à (e_{12}, \dots, e_{1K}) , et désignons par $v_l^2 = (f_{11} - \tilde{f}_l' F_{22}^{-1} \tilde{f}_l)$ la variance des résidus de la régression.

L'équation (2.12) découle directement de (A2) puisque

$$\tilde{f}_l' F_{22}^{-1} = -\frac{1}{v_l^2} \tilde{h}_l'; \quad (f_{11} - \tilde{f}_l' F_{22}^{-1} \tilde{f}_l)^{-1} = h_{11} \quad (A3)$$

d'après les propriétés de l'inverse d'une matrice partitionnée.

b) Détermination de l'équation (4.4)

D'après (4.3),

$$\tilde{\Theta}_l^{(A)} = (I - K_{(P)}' Z_{(A)}') T \tilde{\Theta}_l^{(A)-1} + K_{(P)}' \tilde{Y}_{(A)} \quad (A4)$$

Par conséquent,

$$\tilde{\Theta}_l^{(A)} - \tilde{\Theta}_l = (I - K_{(P)}' Z_{(A)}') (T \tilde{\Theta}_l^{(A)-1} - \tilde{\Theta}_l) + K_{(P)}' (\tilde{Y}_{(A)} - Z_{(A)}' \tilde{\Theta}_l). \quad (A5)$$

Les erreurs de prédiction $(T \tilde{\Theta}_l^{(A)-1} - \tilde{\Theta}_l)$ sont indépendantes des résidus $(\tilde{Y}_{(A)} - Z_{(A)}' \tilde{\Theta}_l)$; par conséquent,

$$P_{(A)}' = E[(\tilde{\Theta}_l^{(A)} - \tilde{\Theta}_l)(\tilde{\Theta}_l^{(A)} - \tilde{\Theta}_l)'] = \tilde{\Theta}_l P_{(A)}'^{-1} \tilde{\Theta}_l' + K_{(P)}' \Sigma_{(A)} K_{(P)}' \quad (A6)$$

où, pour des raisons de commodité, nous avons posé $\tilde{\Theta}_l = (I - K_{(P)}' Z_{(A)}')$.

La figure 5 montre dans quelle mesure les contraintes linéaires permettent de prévenir les variations subtiles de données; pour cela, nous avons représenté graphiquement les valeurs estimées mensuelles de l'ordonnée à l'origine pour les logements de 3 pièces.

Dans l'hypothèse de contraintes linéaires, l'ordonnée à l'origine réagit aussitôt à une variation brusque des données. En l'absence de contraintes, l'adaptation s'étend sur plusieurs mois. Le même graphique pour les logements de 5 pièces n'est pas aussi simple puisque, compte tenu de la faible taille des échantillons mensuels, le "gonflement" des données s'est répercuté sur les autres coefficients de régression.

Jusqu'à maintenant, notre analyse a porté exclusivement sur la distribution empirique des résidus et des résidus une étape à l'avance du modèle. Or, une application importante de l'estimation pour petites régions est la prévision de moyennes de petites régions (équation 2.2). De toute évidence, lorsqu'un modèle produit des résidus qui ont des propriétés "classiques", on peut s'attendre qu'il produise aussi de bons estimateurs pour les moyennes de population. Néanmoins, il serait intéressant de comparer les variances théoriques des estimateurs de moyenne de petite région calculées, les unes suivant l'hypothèse de la corrélation transversale et les autres suivant l'hypothèse de l'absence de corrélation transversale, en vertu du modèle qui tient compte de cette corrélation ($\rho_j \equiv \frac{1}{2}$). Cette comparaison permet de mesurer la perte d'efficacité subie lorsqu'on fait abstraction de l'autocorrélation.

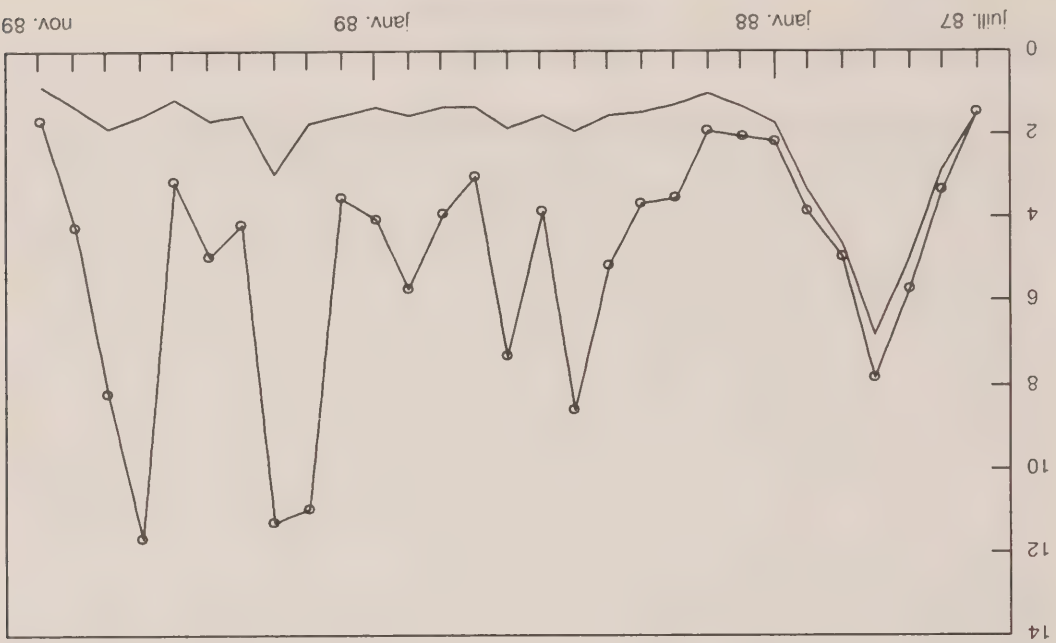
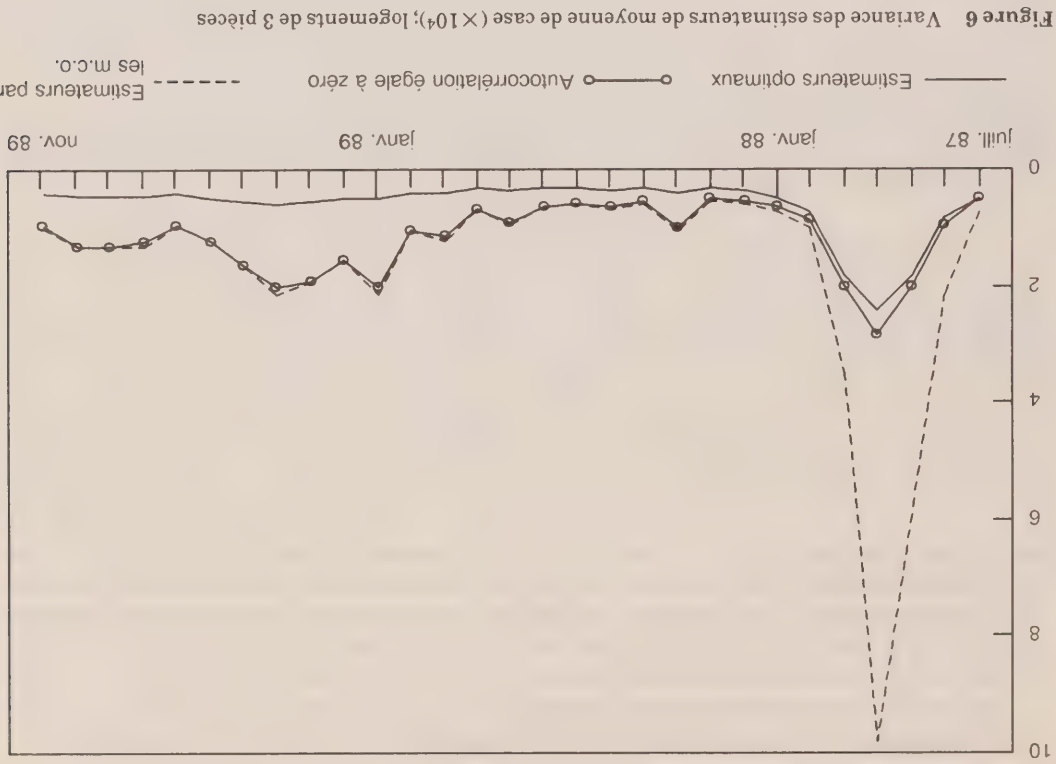
Les figures 6 et 7 présentent les variances mensuelles des estimateurs de moyenne de case pour les logements de 3 et de 5 pièces. (Les variances ont été multipliées par 10^4 .) La figure pour les logements de 3 pièces présente, en plus, les variances des estimateurs par les moindres carrés ordinaires (m.c.o.) de la moyenne de population, c'est-à-dire les variances des estimateurs obtenus lorsqu'on estime les coefficients de régression à chaque mois par les m.c.o. Ce genre d'estimateurs n'est pas vraiment utile pour les logements de 5 pièces à cause de la trop petite taille des échantillons mensuels.

Ce qu'il faut retenir surtout de ces deux graphiques, c'est qu'en tenant compte de la corrélation transversale, on peut réduire de façon substantielle (selon la taille de l'échantillon) la variance des estimateurs. On le voit clairement en ce qui a trait aux logements de 5 pièces; cette constatation vaut aussi pour les logements de 3 pièces même si la taille des échantillons dans ce cas est relativement élevée. D'ailleurs, à cause de la taille ordinairement élevée des échantillons de logements de 3 pièces, les estimateurs par les m.c.o. se rapprochent sensiblement des estimateurs que l'on utilise pour estimer les moyennes de population en faisant abstraction de la corrélation transversale. Notons toutefois l'écart appréciable qui sépare la variance de l'estimateur par les m.c.o. de celle des deux autres estimateurs en octobre 1987. Ce mois-là, il n'y a eu que 10 observations pour les logements de 3 pièces; ce cas illustre très bien l'utilité de recourir à des données de périodes antérieures même lorsqu'on fait abstraction de la corrélation transversale. (Le nombre d'observations pour novembre 1987 est de 28; pour tous les autres mois, on compte au moins 46 observations.)

Les deux graphiques nous amènent à une autre conclusion importante: la variance des estimateurs optimaux selon le modèle est beaucoup plus stable que celle des estimateurs qui font abstraction de la corrélation transversale. Il convient de souligner à cet égard que les différences de variances d'un mois à l'autre ne dépendent pas uniquement de la taille des échantillons mensuels mais aussi de la valeur des variables explicatives (matrice de plan) et de la quantité d'observations de périodes antérieures. C'est néanmoins la taille des échantillons qui explique en majeure partie les différences de variance entre les estimateurs, particulièrement vers la fin de la série.

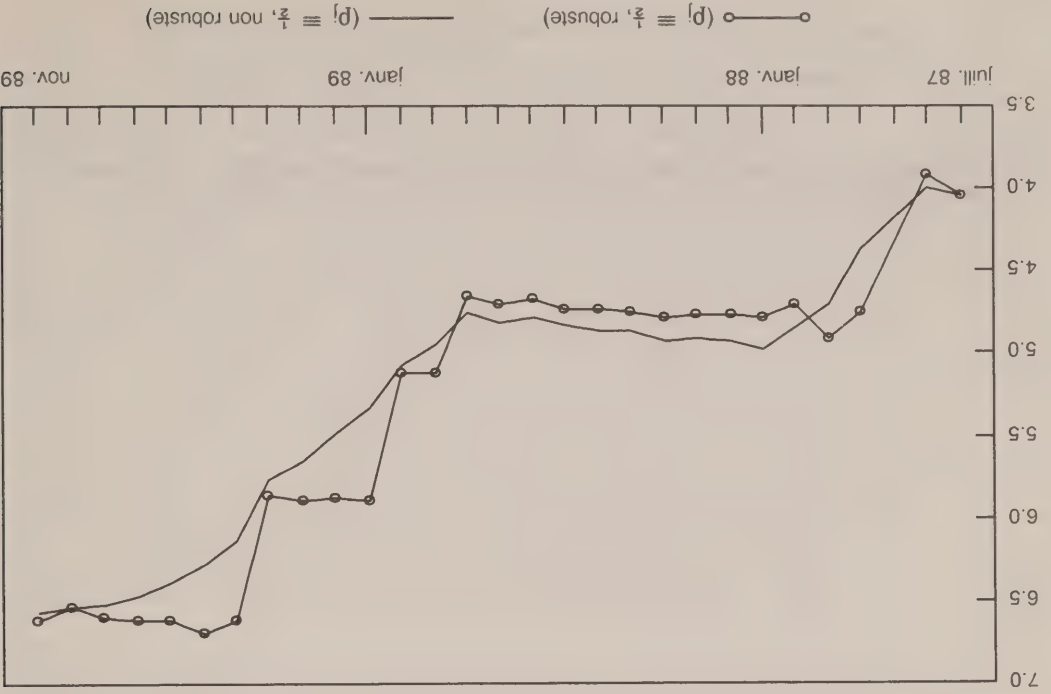
REMERCIEMENTS

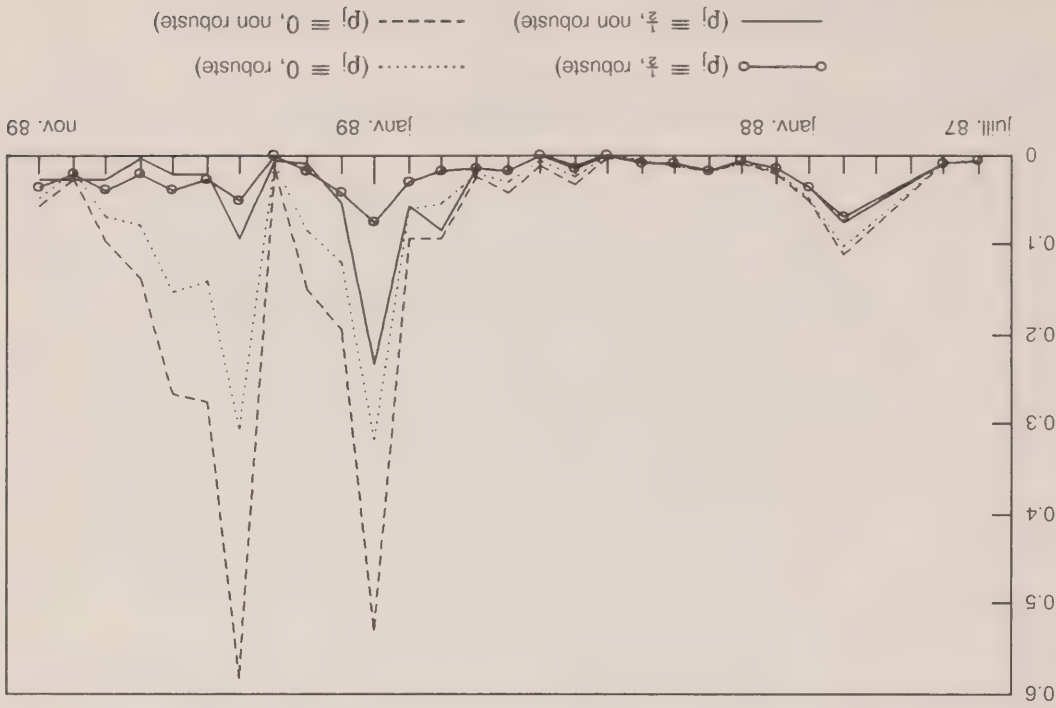
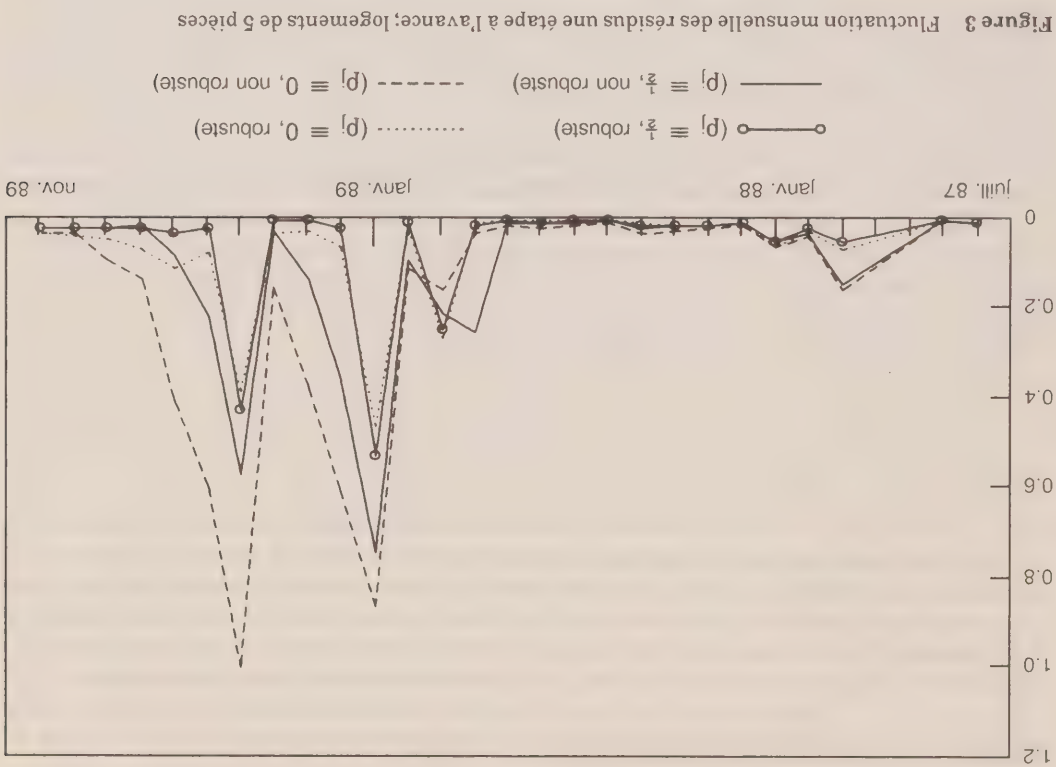
Cet article a été rédigé pendant que le premier auteur était en congé sabbatique à Statistique Canada en vertu du programme de bourses de recherche de cet organisme. Les auteurs tiennent à exprimer leur reconnaissance à l'arbitre pour ses commentaires utiles.



($\rho \equiv \frac{1}{2}$, est. non robuste) représentent leur forme d'avant gonflement beaucoup plus rapidement que celles qui se rapportent au scénario ($\rho \equiv 0$, est. non robuste).

Une autre comparaison intéressante est celle que l'on peut faire entre le cas où on a introduit des contraintes linéaires et celui où on n'en a pas introduites. De toute évidence, l'introduction de contraintes a un effet très favorable lorsqu'on tient compte de l'autocorrélation et cet effet est encore plus favorable lorsque l'autocorrélation est égale à zéro. À cet égard, il est intéressant de comparer les figures qui présentent la fluctuation mensuelle des résidus une étape à l'avance avec celles qui présentent la fluctuation mensuelle des résidus ordinaires. Dans les quatre mois où il y a eu "gonflement" des données, la fluctuation des résidus une étape à l'avance est forte, phénomène parfaitement normal étant donné que ces résidus sont la différence entre les observations et les prédicteurs correspondants des mois précédents. Or, lorsqu'on fait intervenir les contraintes linéaires, la fluctuation revient à son niveau normal dans le mois qui suit le mois du "gonflement". Pour ce qui a trait aux résidus ordinaires, dans l'hypothèse de contraintes linéaires, la fluctuation n'est pas vraiment plus forte dans les mois de "gonflement" pour les logements de 3 pièces et à peine plus forte, si l'on tient compte de l'autocorrélation, pour les logements de 5 pièces. Toutefois, si l'on fait abstraction de l'autocorrélation dans ce dernier cas, la fluctuation des résidus est beaucoup plus forte dans les mois de "gonflement" que dans les autres mois, même lorsqu'on fait intervenir les contraintes. Cela s'explique facilement. En effet, nous savons que les contraintes linéaires s'appliquent aux moyennes globales des valeurs ajustées dans chaque district; or, comme le nombre d'observations pour les logements de 5 pièces ne représente qu'une petite fraction du nombre total d'observations, les contraintes influent relativement peu sur les coefficients de régression estimés pour cette catégorie de logements. En revanche, ces mêmes contraintes ont un effet notable sur les coefficients estimés pour les autres catégories de logement, de sorte que si l'on tient compte de la corrélation transversale, les estimateurs relatifs aux logements de 5 pièces subissent quand même des modifications puisqu'ils sont en corrélation avec les autres coefficients.





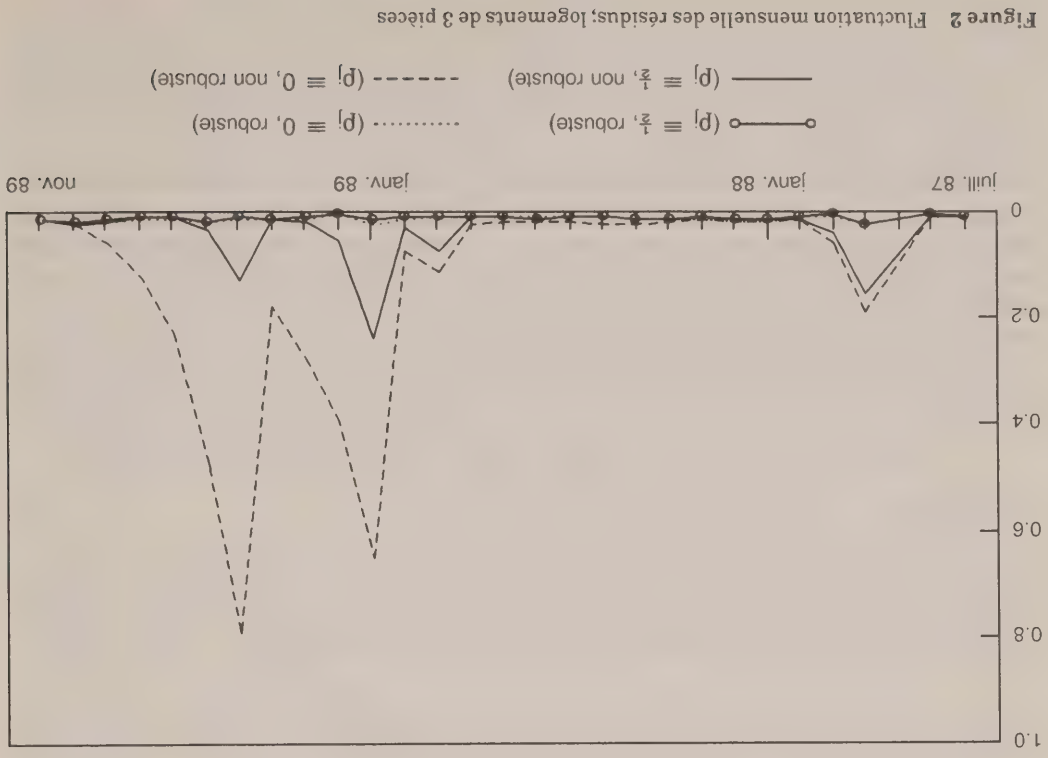
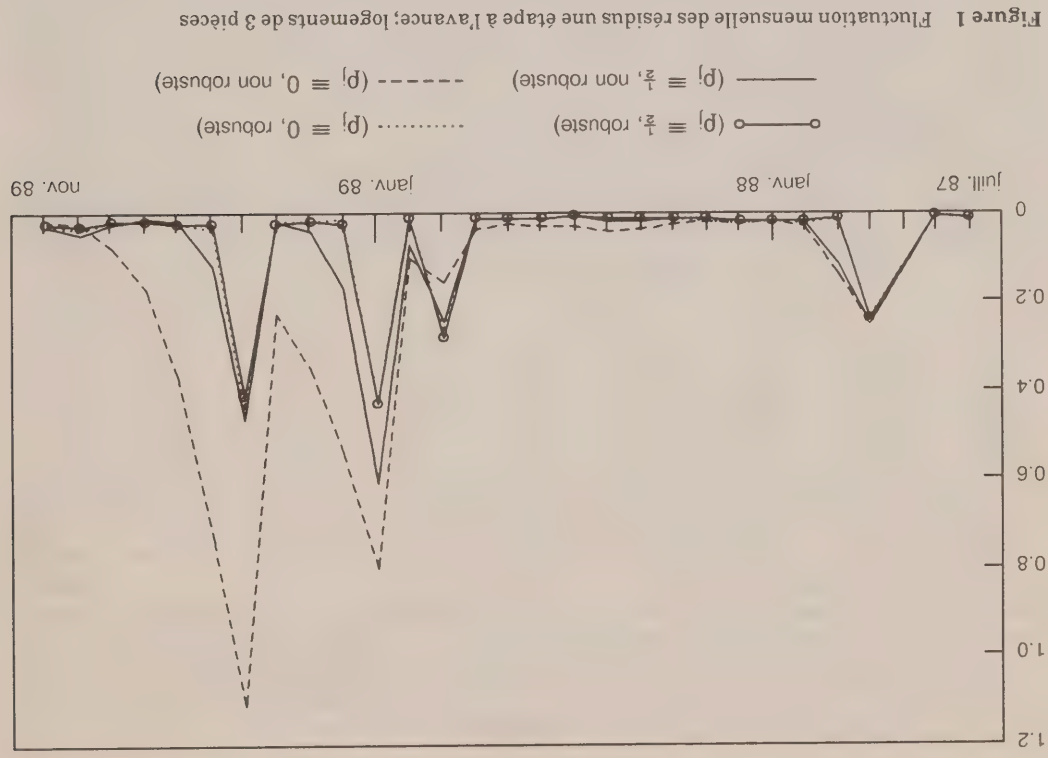


Tableau 1

Fluctuation moyenne des résidus et des résidus une étape à l'avance, compte tenu ou non de la corrélation transversale et du caractère de robustesse; selon la catégorie de logement

Cat. de logement (n° de pièces)	Fluctuation des résidus une étape à l'avance		Fluctuation des résidus	
	Rob.	Non rob.	Rob.	Non rob.
$p \equiv 1/2$				
$p \equiv 0$				
1	.141	.134	.021	.056
2	.070	.090	.021	.023
3	.065	.090	.017	.019
4	.067	.123	.019	.021
5	.067	.114	.023	.065

Le tableau 1 donne la fluctuation moyenne (FM) des résidus $\epsilon_{tki} = (X_{tki} - \beta_{tko} - \sum_{j=1}^4 X_{tkj}^{(j)} \beta_{t-1,kj})$ et des résidus une étape à l'avance $e_{tki} = [X_{tki} - (\beta_{t-1,k0} + \beta_{k0}) - \sum_{j=1}^4 X_{tkj}^{(j)} \beta_{t-1,kj}]$ du modèle (voir équations 5.1 et 5.2) pour chacune des catégories de logement. Les formules utilisées pour calculer cette fluctuation sont $FM_k(\epsilon) = 1/N \sum_{i=1}^N \epsilon_{tki}^2$; $FM_k(e) = 1/N \sum_{i=1}^N e_{tki}^2$ ($1/n_t \sum_{i=1}^n e_{tki}^2$), où $t = 1, \dots, N$ désigne les mois compris entre juillet 1987 et novembre 1989. Nous définissons quatre estimateurs des coefficients de régression, selon que le modèle tient compte ($p_j \equiv 1/2$) ou non ($p_j \equiv 0$) de la corrélation transversale et selon qu'on a modifié ou non les estimateurs de manière à prévenir les défaillances du modèle (dans le tableau, ce dernier critère est représenté par les rubriques "robuste", "robust" et "non robuste" (non rob.)). Pour modifier les estimateurs, nous avons ajouté à l'équation d'observation de chaque mois trois contraintes linéaires comme celle définie en (4.2). Ces contraintes ont pour effet de faire concorder la moyenne globale des valeurs ajustées, dans chacun des trois districts, avec la moyenne correspondante des valeurs observées. Lorsque nous avons introduit les contraintes, nous avons ajusté le modèle à l'aide du filtre de Kalman modifié défini par les équations (4.3) et (4.4).

Afin d'illustrer l'efficacité des quatre séries d'estimateurs par régression aux divers mois et plus particulièrement autour des mois où nous avons "gonflé" les observations, nous avons représenté graphiquement la fluctuation mensuelle des résidus une étape à l'avance et des résidus ordinaires pour les logements de 3 et de 5 pièces. Les graphiques pertinents sont reproduits dans les figures 1 à 4. Notons que les valeurs indiquées dans le tableau 1 pour les logements de 3 et de 5 pièces sont les moyennes respectives des valeurs reproduites dans les quatre graphiques. Le tableau et les graphiques nous permettent de tirer les conclusions suivantes:

Lorsqu'on tient compte de la corrélation transversale et qu'on introduit les contraintes linéaires visant à rendre robuste l'estimateur, on obtient de meilleurs résultats que dans tous les autres cas. Cela est particulièrement notable dans le cas des logements d'une pièce et de cinq pièces, pour lesquels il y a très peu d'observations à chaque mois. En ce qui a trait aux trois autres catégories de logement, on n'observe que de faibles différences entre le scénario ($p \equiv 1/2$, est. robuste) et le scénario ($p \equiv 0$, est. robuste), ce qui était prévisible étant donné que la quantité d'information empruntée aux cases avoisinantes (dans un contexte plus général: petites régions) va en diminuant lorsque le nombre d'observations mensuelles augmente. Toutefois, la situation est tout autre lorsqu'on fait abstraction des contraintes linéaires. Dans un tel cas, on obtient de bien meilleurs résultats lorsqu'on tient compte de la corrélation transversale que lorsqu'on n'en tient pas compte et ce, pour toutes les catégories de logement. Ainsi, grâce à l'emprunt d'information, les estimateurs des coefficients de régression s'adaptent beaucoup plus rapidement à une variation subite des données, comme le montrent clairement les figures 1 à 4. Les quatre sommets de chaque graphique correspondent aux mois où il y a eu un "gonflement" des données et comme nous pouvons le voir, les courbes se rapportant au scénario

(voir exemple (b) dans la section 2.1). Ainsi, $\tilde{\alpha}'_{ik} = (\beta_{ik0}, \beta_{ik1}, \dots, \beta_{ik4})$, $Z_{ik} = [\tilde{1}_{nik}, \tilde{0}_{nik}, \tilde{X}_{ik}^{(1)}, \dots, \tilde{X}_{ik}^{(4)}]$, $T = [\tilde{e}_1, \tilde{e}_1 + \tilde{e}_2, \tilde{e}_3, \dots, \tilde{e}_6]$, une matrice 6×6 où \tilde{e}_j renferme le nombre un à la position j et des zéros aux autres positions, et $\tilde{G} = [\tilde{e}_1, \tilde{e}_3, \dots, \tilde{e}_6]$, une matrice 6×5 . La matrice Δ est définie comme en (2.5). Le vecteur $\tilde{\alpha}_i$ et les matrices Z_i , T , G et Δ sont déterminées à l'aide des vecteurs $\{\tilde{\alpha}_{ik}\}$ et des matrices $\{Z_{ik}\}$, T , G et Δ de la même manière qu'en (2.7) et (2.8).

Après avoir exprimé le modèle sous forme de modèle d'états, nous avons cherché à estimer les variances et les covariances inconnues à l'aide de la méthode de la fonction de caractérisation décrite dans la section 3.2. Or, nous nous sommes rendu compte que la puissance de l'ordinateur qui a servi à cette étude, un IBM 1481, était de beaucoup inférieure à ce qu'il aurait fallu avoir pour obtenir la convergence dans des conditions raisonnables. Il convient de souligner que le modèle d'états combiné renferme vingt-cinq paramètres inconnus ($\dim(\tilde{\Delta}) = 25$) tandis que les vecteurs d'états et les matrices de variances-covariances de ces vecteurs ont une dimension de 30 ($\dim(\tilde{\alpha}_i') = 30$). Le nombre mensuel d'observations varie de 55 à 353. Le programme qui a été élaboré pour cette étude utilise des différences finies, de sorte qu'à chaque itération de la méthode de la fonction de caractérisation, il doit se faire une exploration intégrale des données, chaque exploration comportant $[\dim(\tilde{\Delta}) + 1]$ opérations de calcul pour le vecteur d'états $\tilde{\alpha}_i$ et la matrice de variances-covariances P_i (équation 3.1) pour chaque période. Ces calculs sont nécessaires pour évaluer les fonctions de vraisemblance logarithmiques et les différences finies correspondantes. C'est donc dire que les frais de calcul augmentent avec le nombre d'observations, la dimension du vecteur d'états et le nombre de paramètres inconnus.

Pour contourner ce problème, nous avons estimé séparément la variance σ_k^2 (équation 2.1) et la matrice Δ (équation 2.5) pour chacune des catégories de logement au moyen des séries d'observations respectives, puis avons estimé les coefficients de corrélation ρ_j (équation 2.6) par une simple recherche par quadrillage. Nous avons constaté que le fait de poser ρ_j égal à $\frac{1}{2}$, pour chaque j , donnait des résultats satisfaisants tant au point de vue de la tendance des résidus qu'une étape à l'avance (erreurs de prédiction une étape à l'avance) qu'au point de vue du lissage des coefficients de régression relatifs aux logements d'une pièce et de cinq pièces, pour lesquels il y a très peu d'observations à chaque mois. Précisons qu'en estimant séparément les variances et les covariances qui définissent la corrélation longitudinale des coefficients de régression propres à chaque catégorie, on déroge quelque peu aux hypothèses du modèle, mais il faut dire qu'il y a de toute évidence une perte d'efficacité si les variances et les covariances sont identiques d'une catégorie à l'autre.

5.3 Résultats

Pfeiffermann, Burck et Ben-Tuvia (1989) montrent jusqu'à quel point les modèles chronologiques peuvent être ajustés convenablement aux données relatives aux diverses catégories de logement. Dans la présente étude, nous cherchons plutôt à comparer les résultats obtenus lorsqu'on tient compte de la corrélation transversale avec ceux obtenus lorsqu'on n'en tient pas compte, et à montrer dans quelle mesure les modifications proposées en (4.1) permettent de prévenir les défaillances du modèle.

Afin de rendre la comparaison la plus claire possible, nous avons augmenté délibérément les valeurs de Y de 5% à chacune des périodes: octobre 1987, novembre 1988, janvier 1989 et mai 1989. Ainsi, toutes les valeurs de Y enregistrées de octobre 1987 à octobre 1988 pour les cinq catégories de logement ont été 'gonflées' de 5%, les valeurs de Y enregistrées en novembre et en décembre 1988 ont été haussées de 10.25% (c'est-à-dire 5% ajouté au 5% précédent) et ainsi de suite. Ce genre de défaillances peut survenir (quoique dans des conditions évidemment beaucoup moins sévères) par suite d'une dévaluation de la monnaie par exemple, et il importe d'en tenir compte dans la modélisation des prix de vente. Voir Pfeiffermann, Burck et Ben-Tuvia (1989) pour une analyse approfondie. Des défaillances semblables peuvent survenir par exemple pour les séries de prix de loyer de chômage dans les périodes de récession profonde.

données pour le calcul des indices mensuels des prix du logement (IPL) corrigés en fonction des changements qualitatifs. On calcule un IPL pour chaque ville ou chaque ensemble de villes et pour chaque catégorie de logement, définie par le nombre de pièces (de 1 à 5). Le nombre mensuel de transactions est très peu élevé pour de nombreuses combinaisons "ville/catégorie de logement" et il arrive qu'il n'y ait aucune transaction dans le cas des logements d'une pièce. Le tableau ci-dessous donne le nombre mensuel moyen de transactions enregistrées entre juillet 1987 et novembre 1989, ainsi que l'écart-type (E.T.) correspondant.

Catégorie	1	2	3	4	5
Moyenne	2.7	29.0	101.9	39.7	5.6
E.T.	2.6	12.9	50.4	18.8	3.5

La nécessité de tenir compte des changements qualitatifs tient à ce que les transactions ne sont soumises à aucun contrôle, ce qui engendre de fortes variations qualitatives d'un mois à l'autre, particulièrement dans les cas à faible fréquence. Pour chaque transaction, on enregistre les variables de mesure de la qualité (VMQ) suivantes: $X^{(1)}$ - superficie du logement; $X^{(2)}$ - âge du logement; $X^{(3)}$, $X^{(4)}$ - variables auxiliaires servant à désigner les districts de la ville. Dans un article récent, Pfeffermann, Burck et Ben-Tuvia (1989) analysent en détail les problèmes liés au calcul de l'IPL et la méthode de calcul utilisée en Israël. Ils proposent le modèle ci-dessous comme solution de rechange pour le modèle en usage actuellement. L'indice triple "tki" sert à désigner la transaction i dans la catégorie k au mois t , $X^{tki}_{(t)}$ étant le logarithme du prix de vente correspondant et $X^{tki}_{(t)}$ = $\log(X^{tki}_{(t)})$, $j = 1, 2$.

(5.1)
$$Y_{tki} = \beta_{tk0} + \beta_{tk1}X^{tki}_{(1)} + \beta_{tk2}X^{tki}_{(2)} + \beta_{tk3}X^{tki}_{(3)} + \beta_{tk4}X^{tki}_{(4)} + \epsilon_{tki}$$

(5.2)
$$\begin{aligned} \beta_{tk0} &= \beta_{t-1,k0} + \beta_{k0} + \eta_{tk0} \\ \beta_{tkj} &= \beta_{t-1,kj} + \eta_{tkj}, j = 1, \dots, 4, \end{aligned}$$

où les termes d'erreur ϵ_{tki} et η_{tkj} satisfont aux hypothèses (2.1), (2.4) et (2.5). Notons que le modèle posé pour l'origine correspond à la formule d'approximation locale d'une tendance linéaire, définie dans l'exemple (d) de la section (2.1). Le modèle posé pour les autres coefficients correspond au modèle de trajet aléatoire défini dans l'exemple (b). À partir du modèle de régression défini en (5.1), on peut construire un IPL corrigé en fonction des changements qualitatifs. En prenant pour valeur des VMQ la moyenne de population correspondante, qui est constante en pratique (la valeur de ces variables étant révisée à tous les cinq ans environ), on peut calculer des prix de vente moyens à l'aide de (5.1) et ces moyennes sont comparables d'un mois à l'autre puisqu'elles se rapportent à des logements de même nature. Pfeffermann, Burck et Ben-Tuvia examinent les raisons qui motivent le choix du modèle (5.2) pour les coefficients de régression. Ils présentent aussi des résultats empiriques qui confirment la validité du modèle. Seulement, ces résultats ont été obtenus en ajustant le modèle aux données de chaque case prise individuellement, c'est-à-dire sans tenir compte de la corrélation transversale entre les coefficients de régression. Dans la présente étude, nous approfondissons cet aspect du modèle et, chose non moins importante, nous donnons un aperçu de l'efficacité des modifications proposées dans la section 4 pour prévenir les défaillances du modèle.

5.2 Estimation du modèle

Le modèle défini en (5.1) et (5.2) peut être exprimé sous forme de modèle d'espace d'états comme dans les équations (2.7) et (2.8). En fait, les vecteurs g_t et les matrices Z_t , T et G impliquent en l'occurrence des structures simples puisque pour $j = 1, \dots, 4$, $\beta_{kj} \equiv 0$

correspondants $(W_{11}^{(0)} I_{n_1} Z_{11}', \dots, W_{1K}^{(0)} I_{n_K} Z_{1K}')$ et de poser les variances des résidus égales à zéro. Le nouveau système d'équations, y compris (2.8), constitue un pseudo-modèle d'espace d'états qui peut être estimé à l'aide des équations du filtre de Kalman (3.1). Notons que la pseudo-matrice de variances-covariances $\Sigma_{(P)}'$ du vecteur élargi des résidus n'est plus définie positive (les $L(t)$ dernières lignes et $L(t)$ dernières colonnes de $\Sigma_{(P)}'$ sont formées de zéros) mais cela ne complique pas le calcul.

L'application du filtre de Kalman au pseudo-modèle présente un inconvénient; en effet, les matrices de variances-covariances des estimateurs par régression ne tiennent pas compte de la variabilité réelle des moyennes globales dans le membre de gauche de l'équation (4.1). Pour remédier à cela, nous proposons de modifier la formule de mise à jour de la matrice de variances-covariances P_t (équation 3.1) de manière que l'on tienne compte des variances et des covariances des moyennes globales.

Désignons par $\tilde{Y}_{(A)}'$ et $Z_{(A)}'$ le vecteur Y et la matrice Z élargis au temps t et par $\Sigma_{(A)}'$ la vraie matrice de variances-covariances des résidus $[\tilde{Y}_{(A)}' - Z_{(A)}' \tilde{q}_t']$. La matrice $\Sigma_{(A)}'$ est d'ordre $[n_t + L(t)]$, les éléments de $\Sigma_{(A)}'$ occupant les n_t premières lignes et n_t premières colonnes, avec les variances et les covariances des moyennes $\Sigma_K W_{(0)}^{(K)} \Sigma_t Y^{(K)t}$ et les éléments du vecteur \tilde{Y}_t occupant le reste des lignes et des colonnes. En désignant par $\tilde{q}_{(A)}'$ le prédicteur robuste de \tilde{q}_{t-1} déterminé à $(t-1)$ à l'aide du pseudo-modèle et par $P_{(A)}'^{-1}$, la vraie matrice de variances-covariances des erreurs $(\tilde{q}_{(A)}'^{-1})$, nous pouvons déterminer l'estimateur d'état modifié au temps t par la formule

$$\tilde{q}_{(A)}' = T \tilde{q}_{(A)}'^{-1} + P_{(A)}'^{-1} Z_{(A)}' (F_{(P)}')^{-1} [\tilde{Y}_{(A)}' - Z_{(A)}' T \tilde{q}_{(A)}'^{-1}] \quad (4.3)$$

où $P_{(A)}'^{-1} = (T P_{(A)}'^{-1} T' + G A G')$ et $F_{(P)}' = Z_{(A)}' P_{(A)}'^{-1} Z_{(A)}' + \Sigma_{(P)}'$ (comparer avec les équations (3.1)). Nous montrons en annexe que la matrice réelle des variances-covariances $(P_{(A)}')$ des erreurs $(\tilde{q}_{(A)}')$ satisfait l'équation récurrente

$$P_{(A)}' = [I - K_{(P)}' Z_{(A)}' P_{(A)}'^{-1} + K_{(P)}' [\Sigma_{(A)}' - \Sigma_{(P)}'] K_{(P)}'] \quad (4.4)$$

où $K_{(P)}' = P_{(A)}'^{-1} Z_{(A)}' (F_{(P)}')^{-1}$ est le pseudo-gain de Kalman. Le premier terme du membre de droite de l'équation (4.4) représente la formule de mise à jour usuelle du filtre de Kalman (comparer avec (3.1)). Le second terme est un facteur de correction qui, tient compte des variances et des covariances réelles des moyennes $\Sigma_K W_{(0)}^{(K)} \Sigma_t Y^{(K)t}$, qu'ignore le premier terme.

Le filtre de Kalman modifié défini en (4.3) et (4.4) produit le prédicteur robuste $\tilde{q}_{(A)}'$ au lieu du prédicteur optimal (fondé sur un modèle) \hat{q}_t , mais utilise la bonne matrice de variances-covariances suivant le modèle. Il peut donc servir régulièrement à estimer les vecteurs de coefficients et, par voie de conséquence, les moyennes de petites régions; de plus, lorsque le modèle est valide, il donnera des résultats comparables à ceux obtenus avec le filtre optimal. Dans les cas où le modèle ne serait pas valide, l'équation (4.4) pourrait être incorrecte (suivant le genre de défaillance qui affecte le modèle) mais les prédicteurs $\tilde{q}_{(A)}'$ satisferont quand même les contraintes linéaires (4.1). De même, il est possible de modifier les équations de lissage (3.2) pour qu'elles satisfassent les contraintes linéaires.

5. RÉSULTATS EMPIRIQUES

5.1 Description des données et du modèle ajusté

Afin d'illustrer les grandes caractéristiques de la classe de modèles définis dans la section 2, nous avons ajusté ces modèles aux prix de vente des maisons à Jérusalem. Le Bureau central de la statistique d'Israël fait un relevé de ces prix à chaque mois et utilise régulièrement ces

On peut, par exemple, modifier les estimateurs par régression calculés aux diverses périodes de manière qu'ils satisfassent certaines contraintes linéaires que l'on aura établies en posant la moyenne globale des données brutes égale à sa valeur ajustée probable selon le modèle. Plus précisément, nous proposons d'ajouter à l'équation de modèle (2.1) des contraintes linéaires du type

$$\sum_k W_{ik}^{(t)} \sum_i Y_{ik} = \sum_k W_{ik}^{(t)} \sum_i \hat{x}_{ik} \hat{\theta}_{ik} \quad i = 1, 2, \dots, L(t), \quad t = 1, \dots, t^* \quad (4.1)$$

où les coefficients $W_{ik}^{(t)}$ sont des poids normalisés fixes tels que $\sum_k n_{ik} W_{ik}^{(t)} = 1$. Un exemple de contrainte linéaire est l'équation

$$\sum_K N_{ik} \hat{M}_{ik} \left/ \sum_K N_{ik} \right/ = \sum_K N_{ik} (\hat{x}_{ik} \hat{\theta}_{ik}) \left/ \sum_K N_{ik} \right/ \quad (4.2)$$

où M_{ik} est l'estimateur d'enquête (direct) pour la région k . Lorsque $\hat{x}_{ik} \approx \hat{X}_{ik}$, l'équation (4.2) fait que le prédicteur de la moyenne globale de population (fondé sur un modèle) concorde avec l'estimateur d'enquête correspondant. On peut justifier cette contrainte en faisant valoir que, même s'ils ne sont pas assez fiables pour l'estimation des moyennes de petites régions, vu la faible taille des échantillons, les estimateurs d'enquête peuvent être jugés très satisfaisants lorsqu'ils servent collectivement à estimer la moyenne globale. Notons, de plus, que les organismes statistiques doivent de toute façon définir des contraintes pour les besoins de leurs analyses. Battese, Harter et Fuller (1988) et Pfeffermann et Barnard (1991) utilisent une contrainte semblable pour analyser des enquêtes transversales. Il arrive souvent que l'on puisse former des groupes avec les petites régions, à la condition qu'il y ait suffisamment de données dans chacun des groupes pour pouvoir estimer les moyennes de groupes à l'aide des estimateurs d'enquête. On peut alors introduire plusieurs contraintes comme celle définie en (4.2), la sommation s'étendant cette fois aux régions du même groupe. À cet égard, il convient de souligner que, compte tenu de la corrélation entre les coefficients de régression dans les diverses régions, l'application d'une contrainte à un sous-ensemble des régions influera sur les estimations de toutes les régions. Un exemple illustrera cette propriété.

Il importe de souligner que la série de contraintes définie en (4.1) ne représente pas de l'information externe sur les valeurs probables des coefficients de régression. Elle constitue plutôt un "système de contrôle" qui vise à ce que les estimateurs du modèle réagissent plus rapidement à une variation éventuelle des coefficients de régression. C'est pourquoi les variances des estimateurs par régression modifiées sont légèrement plus élevées que celles des estimateurs optimaux. Évidemment, lorsqu'il ne se produit aucun changement et que les variances des moyennes globales sont suffisamment faibles, on s'attend que les contraintes soient à peu près satisfaites sans même qu'elles soient introduites explicitement. Comme nous l'avons mentionné plus haut, il est possible d'introduire plusieurs contraintes différentes à chaque période mais il faut absolument que les variances des moyennes globales correspondantes soient suffisamment faibles pour que les modifications soient vraiment nécessaires et qu'elles ne se confondent pas avec les fluctuations aléatoires des données brutes.

4.2 Inférence avec contraintes linéaires

Dans la section 4.1, nous avons proposé de modifier l'équation de modèle (2.1) en y ajoutant la série de contraintes (4.1) de manière que les estimateurs par régression résistent bien à une variation subite des valeurs des coefficients.

La meilleure façon de procéder sur le plan du calcul est d'ajouter au vecteur \hat{Y}_i de l'équation (2.7) la grandeur scalaire $\sum_k W_{ik}^{(t)} \sum_i Y_{ik}$, d'ajouter aux matrices Z_i les vecteurs lignes

Il faut initialiser le filtre de Kalman pour calculer la fonction de vraisemblance; la meilleure façon d'exécuter cette opération est de recourir à la méthode proposée par Harvey et Phillips (1979). Selon cette méthode, les éléments non stationnaires du vecteur d'états sont initialisés avec une variance de l'erreur très élevée (ce qui équivaut à poser par hypothèse une distribution a priori non informative), de telle sorte que l'on puisse considérer comme nulles les estimations d'états correspondantes. (Pour ce qui a trait au modèle de trajet aléatoire, le fait d'initialiser avec une distribution a priori non informative permet d'établir les estimateurs par les m.c.o. au bout d'une période; voir Meinhold et Singpurwalla 1983, pour une formulation bayésienne du filtre de Kalman.) Quant aux éléments stationnaires du vecteur d'états, ils sont initialisés à l'aide des moyennes et des variances inconditionnelles correspondantes, qui peuvent compter parmi les paramètres inconnus qui définissent les arguments de la fonction de vraisemblance. On peut maximiser la fonction de vraisemblance (3.4) par la méthode de la fonction de caractérisation avec pas variable. Plus particulièrement, désignons par $\hat{\lambda}^{(o)}$ les valeurs estimées initiales des éléments inconnus dans $\hat{\lambda}$. La méthode de la fonction de caractérisation consiste alors à résoudre itérativement le système d'équations

$$\hat{\lambda}^{(i)} = \hat{\lambda}^{(i-1)} + r_i \{ I[\hat{\lambda}^{(i-1)}] \}^{-1} g[\hat{\lambda}^{(i-1)}] \quad (3.5)$$

où $\hat{\lambda}^{(i-1)}$ est l'estimateur de $\hat{\lambda}$ obtenu à la $(i - 1)$ ième itération, $I[\hat{\lambda}^{(i-1)}]$ est la matrice d'information calculée en fonction de $\hat{\lambda}^{(i-1)}$ et $g[\hat{\lambda}^{(i-1)}]$ est le gradient de la fonction de vraisemblance logarithmique calculé à $\hat{\lambda}^{(i-1)}$. Le coefficient r_i désigne un pas variable qui fait que la condition $L[\hat{\lambda}^{(i)}] \geq L[\hat{\lambda}^{(i-1)}]$ est respectée à chaque itération. La valeur de ce coefficient peut être déterminée au moyen d'une recherche par quadrillage dans la région $[0, 1]$. Les formules pour l'élément k du vecteur gradient et l'élément k, l de la matrice d'information sont définies dans Watson et Engle (1983). Une fois que les variances et les covariances du modèle ont été estimées, on peut substituer ces estimations aux paramètres réels des équations du filtre de Kalman (équ. 3.1 et 3.2) de manière à obtenir les estimateurs des coefficients de régression et des matrices de variances-covariances et, de là, les estimateurs pour petites régions et les variances-covariances correspondantes (voir équation 3.3). Soulignons toutefois que les matrices de variances-covariances estimées ne tiennent pas compte de la variation qui découle de la nécessité d'estimer les éléments inconnus compris dans $\hat{\lambda}$. Pour tenir compte de cette variation additionnelle dans l'élaboration de modèles d'espace d'états, Ansley et Kohn (1986) proposent l'utilisation de facteurs de correction d'ordre $1/1^*$ en ayant recours à des formules d'approximation de Taylor du premier ordre. Pour sa part, Hamilton (1986) propose une méthode du type Monte Carlo qui consiste à échantillonner des unités provenant d'une distribution normale multidimensionnelle dont la moyenne est définie par l'estimateur du maximum de vraisemblance du vecteur $\hat{\lambda}$ et la matrice de variances-covariances, par l'inverse de la matrice d'information, et à estimer les vecteurs d'états pour chaque réalisation aléatoire des valeurs de paramètres. Cette méthode offre plus de souplesse au point de vue des hypothèses et nous renseigne davantage sur la sensibilité des estimateurs du filtre de Kalman par rapport à l'erreur dans les estimateurs de la variance et de la covariance. En revanche, elle est plus exigeante sur le plan du calcul.

4. MODIFICATIONS VISANT À PRÉVENIR LES DÉFAILLANCES DU MODÈLE

4.1 Description du problème et modifications proposées

L'utilisation d'un modèle paraît inévitable dans l'estimation pour petites régions compte tenu de la faible taille des échantillons. Ceci étant dit, il faut se demander comment prévenir les défaillances du modèle. Tester le modèle à chaque fois que de nouvelles données sont connues est peu pratique; il faut plutôt envisager un "dispositif intégré" qui conserve la robustesse des estimateurs lorsque le modèle n'est pas valide.

où $\tilde{Y}^{t|t-1}_{\hat{q}^{t|t-1}} = Z^{t|t-1}_{\hat{q}^{t|t-1}}$ est le MPLS de \tilde{Y}_t au temps $(t-1)$ de sorte que $\tilde{e}_t = (\tilde{Y}_t - \tilde{Y}^{t|t-1}_{\hat{q}^{t|t-1}})$ est le vecteur des résidus une étape à l'avance ayant comme matrice de variances-covariances $F_t = (Z^{t|t-1}_t Z^{t|t-1}_t + \Sigma_t)$.

Les nouvelles données enregistrées au temps t peuvent servir aussi à la mise à jour (ou au lissage) d'anciens estimateurs des vecteurs d'états et, par voie de conséquence, à la mise à jour d'anciens estimateurs des moyennes de petites régions. En désignant par t^* le dernier mois où ont été enregistrées des observations, on exécute le lissage au moyen des équations

$$\begin{aligned} \hat{q}^{t|t^*}_{\hat{q}^{t|t^*}} &= \hat{q}^t_t + P^t_t T^{t+1|t^*} (\hat{q}^{t+1|t^*}_{\hat{q}^{t+1|t^*}} - T \hat{q}^t_t) \\ P^{t|t^*}_{\hat{q}^{t|t^*}} &= P^t_t + P^t_t T^{t+1|t^*} (P^{t+1|t^*}_{\hat{q}^{t+1|t^*}} - P^{t+1|t^*}_{\hat{q}^{t+1|t^*}} T) P^{t|t^*}_{\hat{q}^{t|t^*}}, \quad t = 2, 3, \dots, t^* \end{aligned} \tag{3.2}$$

où $P^{t|t^*}_{\hat{q}^{t|t^*}}$ est la matrice de variances-covariances des erreurs de prédiction $(\hat{q}^{t|t^*}_{\hat{q}^{t|t^*}} - \hat{q}^t_t)$. Notons que $\hat{q}^{t^*|t^*}_{\hat{q}^{t^*|t^*}} = \hat{q}^{t^*}_{\hat{q}^{t^*|t^*}}$ et $P^{t^*|t^*}_{\hat{q}^{t^*|t^*}} = P_{\hat{q}^{t^*|t^*}}$, ceci définissant les valeurs initiales pour les équations de lissage.

Les estimateurs lissés de \hat{q}_t permettent d'obtenir facilement des estimateurs de moyennes de petites régions ou de moyennes globales à l'aide de la relation $\hat{\Theta}_{tk} = \hat{X}_{tk} \hat{\Theta}_{tk} = \hat{Z}_{tk} \hat{q}_{tk} = \hat{Z}_{tk} A_{tk} \hat{q}_{tk}$ où $\hat{Z}_{tk} = (1, 0, X_{tk1}, 0, \dots, X_{tkp}, 0)$ et A_{tk} est la matrice indicatrice appropriée. Par conséquent, si $\Theta^{t^*}_w = \sum_{k=1}^{K^*} w_k \hat{Z}_{tk} A_{tk} \hat{q}_{tk} = \hat{q}^{t^*w}_{\hat{q}^{t^*|t^*}}$ (par exemple). Pour des matrices de variances-covariances Σ_t et Λ données, l'EQM des erreurs d'estimation est calculée au moyen des équations

$$E(\hat{\Theta}_{tk} - \Theta_{tk})^2 = \tilde{Z}_{tk} A_{tk} P^t_t A_{tk} \tilde{Z}_{tk} \quad \text{et} \quad E(\hat{\Theta}^{t^*}_w - \Theta^{t^*}_w) = \tilde{q}^{t^*w}_{\hat{q}^{t^*|t^*}}. \tag{3.3}$$

Il convient de souligner que l'EQM définie ci-dessus a rapport à la distribution conjointe des observations $\{\tilde{Y}_{tk}\}$ et des vecteurs de coefficients $\{\hat{\Theta}_{tk}\}$ de sorte qu'elle constitue une moyenne pour toutes les valeurs possibles de la moyenne de petite région.

3.2 Estimation des matrices de variances-covariances et initialisation du filtre

Pour appliquer le filtre de Kalman, il est nécessaire d'estimer les éléments inconnus des matrices Σ_t et Λ et d'initialiser le filtre, c'est-à-dire d'estimer le vecteur \hat{q}_0 et la matrice de variances-covariances correspondante (P_0) des erreurs d'estimation. Dans cette section, nous décrivons des méthodes d'estimation simples qui peuvent servir à ces calculs.

Dans l'hypothèse que les résidus \tilde{e}_t et η_t des équations (2.7) et (2.8) sont distribués suivant une loi normale, nous pouvons exprimer la fonction de vraisemblance logarithmique des vecteurs $\tilde{Y}_{m+1}, \dots, \tilde{Y}_{t^*}$, étant donné les m premiers vecteurs $\tilde{Y}_1, \dots, \tilde{Y}_m$, de la façon suivante:

$$L(\tilde{\lambda}) = \text{constante} - \frac{1}{2} \sum_{t^*=m+1}^t (\log |F_t| + \tilde{e}^{t'}_t F^{-1}_t \tilde{e}_t) \tag{3.4}$$

où $\tilde{\lambda}$ renferme les variances et les covariances inconnues du modèle, exprimées sous forme vectorielle. La grandeur scalaire m représente le nombre de périodes nécessaires pour construire des valeurs initiales pour le filtre de Kalman. (En ce qui concerne le modèle de trajet aléatoire exposé dans la section 2.2, $m = 1$, à la condition que l'on dispose de suffisamment de données pour chaque région de manière à pouvoir déterminer les estimateurs par les m.c.o. (moindres carrés ordinaires) des vecteurs de coefficients). L'expression (3.4) découle de la décomposition de l'erreur de prédiction, voir Schweppe (1965) et Harvey (1981) pour plus de détails. Pour des matrices Σ_t et Λ données, on détermine les résidus \tilde{e}_t et les matrices de variances-covariances F_t à l'aide des équations du filtre de Kalman (3.1).

En nous servant des équations du filtre de Kalman définies dans la section 3, nous montrons en annexe que l'estimateur $\hat{\Theta}^{ik}$ de la moyenne de petite région Θ^{ik} (équation 2.2) peut avoir dans ce cas la forme suivante:

$$\hat{\Theta}^{ik} = \hat{x}_{ik}^{ik} \hat{g}_{t-1,k} + \left(1 - \frac{\sigma_k^2}{n_{ik} v_k^2}\right) (\bar{Y}^{ik} - \hat{x}_{ik}^{ik} \hat{g}_{t-1,k}) + \frac{\sigma_k^2}{n_{ik} v_k^2} \sum_{m=1}^{m \neq k} \gamma_{km} (\bar{Y}^{im} - \hat{x}_{im}^{im} \hat{g}_{t-1,m}) \quad (2.12)$$

où $\{\gamma_{km}\}$ représente les coefficients de régression partielle dans la régression de $e_{ik} = (\bar{Y}^{ik} - \hat{x}_{ik}^{ik} \hat{g}_{t-1,k})$ par rapport aux erreurs de prédiction $\{e_{im} = (\bar{Y}^{im} - \hat{x}_{im}^{im} \hat{g}_{t-1,m})\}$ observées dans les autres régions, et v_k^2 est la variance des résidus (inexpliquée) de la régression.

L'estimateur $\hat{\Theta}^{ik}$ renferme trois composantes: un estimateur "synthétique", $\hat{x}_{ik}^{ik} \hat{g}_{t-1,k}$, où $\hat{g}_{t-1,k}$ est le prédicteur optimal de \hat{g}^{ik} fondé sur toutes les observations enregistrées jusqu'à la période $t - 1$ inclusivement; un "facteur de correction", $(\bar{Y}^{ik} - \hat{x}_{ik}^{ik} \hat{g}_{t-1,k})$, fondé sur l'erreur de prédiction pour la région k , et un "facteur d'ajustement" fondé sur les erreurs de prédiction observées pour les autres régions. Les deux premières composantes correspondent à celles de l'estimateur classique pour petite région décrit dans l'introduction. Notons que plus la taille de l'échantillon, n^{ik} , est petite, plus le poids attribué à la moyenne empirique courante \bar{Y}^{ik} pour l'estimation de Θ^{ik} est faible et plus celui attribué au prédicteur chronologique $\hat{x}_{ik}^{ik} \hat{g}_{t-1,k}$ est élevé. Le troisième élément du membre de droite de l'équation (2.12) représente l'information "empruntée" aux régions avoisinantes. Le poids imputé à cet élément dépend du degré de corrélation ρ_j entre les termes d'erreur correspondants $\{\eta_{ikj}\}$ des modèles pour les coefficients de régression (équation 2.11). Evidemment, lorsque les variables explicatives dans les diverses régions sont indépendantes de telle sorte que $\rho_j = 0$ pour tous j , par conséquent, $\gamma_{km} = 0$ pour tous m , le troisième élément disparaît et le prédicteur $\hat{\Theta}^{ik}$ se réduit à une moyenne pondérée de la moyenne courante \bar{Y}^{ik} et du prédicteur chronologique $\hat{x}_{ik}^{ik} \hat{g}_{t-1,k}$.

3. ESTIMATION ET INITIALISATION DE MODELE AU MOYEN DU FILTRE DE KALMAN

3.1 Estimation des coefficients de régression au moyen du filtre de Kalman

Dans cette section, nous présentons les équations du filtre de Kalman servant à la mise à jour et au lissage des vecteurs d'états $\hat{\alpha}_t$ définis par les équations (2.7) et (2.8) (dans le cas qui nous occupe, il s'agit des coefficients de régression de région). Nous supposons que les matrices de variances-covariances Σ_t et Λ sont connues. L'estimation de ces matrices fait l'objet de la section 3.2. La théorie du filtre de Kalman est exposée dans de nombreux ouvrages (voir, par exemple, Anderson et Moore 1979 et Meinhold et Singpurwalla 1983); nous limiterons donc notre analyse aux aspects qui ont le plus rapport à l'estimation pour petits domaines. Soit $\hat{\alpha}_{t-1}$ le meilleur prédicteur linéaire sans biais (MPLS) de α_{t-1} , fondé sur l'ensemble des observations enregistrées jusqu'à la période $(t - 1)$ inclusivement. Comme $\hat{\alpha}_{t-1}$ est MPLS de α_{t-1} , $\hat{\alpha}_{t-1} = T \hat{\alpha}_{t-1}$ est le MPLS de $\hat{\alpha}_t$ au temps $(t - 1)$. De plus, si $P_{t-1} = E(\hat{\alpha}_{t-1} - \alpha_{t-1})(\hat{\alpha}_{t-1} - \alpha_{t-1})'$ est la matrice de variances-covariances des erreurs de prédiction au temps $(t - 1)$, $P_{t|t-1} = TP_{t-1}T' + GAG'$ est la matrice de variances-covariances des erreurs de prédiction ($\hat{\alpha}_{t|t-1} - \alpha_t$) (découle directement de (2.8)).

Lorsqu'un nouveau vecteur d'observations $[\hat{Y}_t, Z_t]$ est connu, le prédicteur de $\hat{\alpha}_t$ et la matrice de variances-covariances P_{t-1} sont mis à jour suivant les formules

$$\hat{\alpha}_t = \hat{\alpha}_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1} (\bar{Y}_t - \hat{Y}_{t|t-1}) \quad (3.1)$$

$$P_t = (I - P_{t|t-1} Z_t' F_t^{-1} Z_t) P_{t|t-1}$$

Avant de terminer cette section, nous allons exprimer le modèle défini en (2.1), (2.5) et (2.6) sous forme de modèle d'espace d'états. Ce type de présentation offre des avantages considérables sur le plan du calcul.

Soit $\tilde{Y}' = (\tilde{Y}'_1, \dots, \tilde{Y}'_K)$ le vecteur de dimension $n_t = \sum_k n_{tk}$ des observations pour toutes les régions au temps t et $\tilde{\epsilon}'_t = (\tilde{\epsilon}'_{t1}, \dots, \tilde{\epsilon}'_{tK})$, les résidus de la régression correspondants. Définissons $Z_{tk} = [\tilde{1}_{n_{tk}}, \tilde{0}_{n_{tk}}, \tilde{x}_{tk1}, \dots, \tilde{x}_{tkp}]$, où $\tilde{0}_{n_{tk}}$ est un vecteur de dimension n_{tk} formé de zéros et \tilde{x}_{tkj} est le vecteur des valeurs de la variable auxiliaire j , $j = 1, \dots, p$. Soit Z'_t la matrice diagonale par blocs formée des matrices Z_{tk} . Cette matrice est d'ordre $n_t \times [K \times 2 \times (p + 1)]$. Définissons aussi $\tilde{\alpha}'_t = (\tilde{\alpha}'_{t1}, \dots, \tilde{\alpha}'_{tkp}, \tilde{\alpha}'_{tk})$, $\tilde{\eta}'_t = (\tilde{\eta}'_1, \dots, \tilde{\eta}'_{tk})$, $\tilde{\Sigma}'_t = \text{Diag} [\sigma^2_1 \tilde{1}_{n_{t1}}, \dots, \sigma^2_K \tilde{1}_{n_{tK}}]$, $T = I^K \otimes \tilde{T}$, et $G = I^K \otimes \tilde{G}$.

En nous servant de ces définitions, nous pouvons réexprimer le modèle défini en (2.1), (2.5) et (2.6) sous la forme concise suivante:

$$\tilde{Y}'_t = Z'_t \tilde{\alpha}'_t + \tilde{\epsilon}'_t; E(\tilde{\epsilon}'_t) = \tilde{0}, E(\tilde{\epsilon}'_t \tilde{\epsilon}'_t') = \tilde{\Sigma}'_t \tag{2.7}$$

$$\tilde{\alpha}'_t = T \tilde{\alpha}'_{t-1} + G \tilde{\eta}'_t; E(\tilde{\eta}'_t) = \tilde{0}, E(\tilde{\eta}'_t \tilde{\eta}'_t') = \Lambda_t \tag{2.8}$$

où $\Lambda = [\Lambda_{kt}]$, $k, t = 1, \dots, K$, $\Lambda_{kt} = \Delta$ lorsque $k = t$ et $\Lambda_{kt} = D(\Delta) \theta$ lorsque $k \neq t$. Les matrices Λ_{kt} sont de dimension $(p + 1) \times (p + 1)$.

Le modèle défini en (2.7) et (2.8) est une représentation classique d'un modèle d'espace d'états; voir, par exemple, Anderson et Moore (1979) et Harvey (1984). Selon la formulation des modèles d'états, (2.7) est l'équation d'observation et (2.8), l'équation d'état, où $\tilde{\alpha}'_t$ désigne le vecteur d'états. L'avantage évident de cette formulation est que l'on peut estimer correctement les vecteurs $\tilde{\alpha}'_t$ et de là, les moyennes de population Θ_{tk} , ainsi que la variance des erreurs d'estimation au moyen du filtre de Kalman. Nous nous arrêtons à l'utilisation de ce filtre dans les sections 3 et 4.

2.2 Estimateurs explicites des moyennes de petites régions

Afin de montrer comment, selon le modèle ci-dessus, les données de périodes antérieures ou de régions avoisinantes servent à "renforcer" les estimateurs pour petites régions, nous examinons le cas où le même vecteur \tilde{x}_{tk} de valeurs de variables auxiliaires s'applique à toutes les unités d'une région donnée à une période donnée. Dans ce cas, nous pouvons exprimer l'équation d'observation en fonction des moyennes empiriques, c.-à-d.

$$Y_{tk} = \tilde{x}_{tk} \tilde{\beta}_{tk} + \epsilon_{tk}; E(\epsilon_{tk}) = 0, E(\epsilon^2_{tk}) = \sigma^2_{tk}/n_{tk}, k = 1, \dots, K. \tag{2.9}$$

Supposons que les coefficients de régression suivent un trajet aléatoire (exemple (b) de l'équation 2.3) tel que

$$\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}; E(\eta_{tkj}) = 0, E(\eta_{tkj} \eta_{tk\ell}) = \delta_{j\ell}, j, \ell = 1, \dots, p \tag{2.10}$$

et pour $k \neq m$,

$$E(\eta_{tkj} \eta_{tmj}) = \delta_{jj} \rho_j; E(\eta_{tkj} \eta_{tm\ell}) = 0, j \neq \ell. \tag{2.11}$$

Le modèle de trajet aléatoire suppose que les coefficients s'éloignent lentement de leur valeur initiale et ne tendent pas naturellement à atteindre une valeur moyenne. Evidemment, pour des résidus η_{tkj} tels que $E(\eta^2_{tkj}) = 0$, les coefficients de régression correspondants ne varient pas dans le temps. Notons aussi que puisque $\tilde{\beta}_{tk} = \tilde{\beta}_{t-1,k} + \eta_{tk}$, le prédicteur de $\tilde{\beta}_{tk}$ au temps $(t - 1)$ est identique au prédicteur $\tilde{\beta}_{t-1,k}$ de $\tilde{\beta}_{t-1,k}$.

(c) $T_j = \begin{bmatrix} \rho_{j1} & 1 \\ 0 & 1 \end{bmatrix}$ suppose la relation autorégressive du premier ordre $(\beta_{kj} - \beta_{kj}) = \rho(\beta_{t-1,kj} - \beta_{kj}) + \eta_{tkj}$, envisagée par Rosenberg (1973a).

(d) $T_j = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ implique que $\beta_{tkj} = \beta_{t-1,kj} + \beta_{kj} + \eta_{tkj}$, ce qui représente une approximation locale d'une tendance linéaire (Kitagawa et Gersch 1984). Le coefficient β_{kj} désigne en l'occurrence une pente fixe.

Il convient de souligner que l'on peut se servir de différentes matrices T_j pour diffé-rents coefficients β_{tkj} . De fait, en posant $\alpha_{tk}^j = (\beta_{tk0}, \beta_{tk1}, \beta_{tk2}, \dots, \beta_{tkp}, \beta_{kp})$; $\bar{T} = \text{diag}[T_0, T_1, \dots, T_p]$, une matrice diagonale par blocs, où T_j désigne le bloc j ; $\bar{G} = I_{p+1} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$, où I_{p+1} est la matrice unité d'ordre $p + 1$ et \otimes désigne le produit de Kronecker, et $\bar{\eta}_{tk} = (\eta_{tk0}, \eta_{tk1}, \dots, \eta_{tkp})$, nous pouvons formuler comme suit le modèle composé valant pour les coefficients $\hat{\beta}_{tk}$:

(2.5)
$$\hat{\alpha}_{tk} = \bar{T}\hat{\alpha}_{t-1,k} + \bar{G}\bar{\eta}_{tk}; \quad E(\bar{\eta}_{tk}) = \bar{0}, E(\bar{\eta}_{tk}\bar{\eta}_{t'-a,k}') = A^a\Delta$$

où A^a égale 1 si $d = 0$ et est nul si $d \neq 0$, et où $\Delta = [\delta_{ij}]$ est définie par les variances et covariances δ_{ij} (équation 2.4).

Le modèle défini en (2.5) décrit la variation longitudinale des coefficients de régression pour une région donnée. Ordinairement, les modèles conçus pour tenir compte de la corrélation trans-versale des moyennes de petites régions renferment des effets aléatoires de petit domaine $\{u_k\}$ qui ne varient pas dans le temps. On peut introduire de tels effets dans le modèle général défini en (2.1) et en (2.3) en posant, par exemple, $\tilde{X}_{tk} = \tilde{X}_{tk}u_k + X_{tk}\beta_{tk} + \epsilon_{tk} = X_{tk}^*\beta_{tk}^* + \epsilon_{tk}$ et $u_k = u_{t-1,k} + \eta_{tk}$, où $\text{var}(\eta_{tk}) = \sigma_{\eta}^2$ et $\text{var}(\eta_{tk}) = 0$ pour $t > 1$ (on peut comparer cette dernière équation avec l'exemple (b) ci-dessus). De plus, en supposant la relation auto-régressive de l'exemple (c) pour la variable en cause et en tenant pour fixes les autres coefficients de régression (exemple (a) avec variance des résidus nulle), on obtient un modèle semblable à celui analysé par Choudhry et Rao (1989) sauf que dans ce dernier cas, les résidus d'obser-vations de l'équation (2.1) peuvent être autocorrélés. Notons que l'équation (2.1) renferme maintenant deux ordonnées à l'origine aléatoires mais le modèle demeure identifiable. Choudhry et Rao supposent que les estimateurs d'enquête sont les seules données disponibles; par conséquent, il faudra recourir aux micro-données pour estimer les autocorrélations. Par ailleurs, il est toujours possible de définir un modèle qui tienne compte des autocorrélations. Pour leur part, Choudhry et Rao supposent un modèle AR(1).

Une façon plus générale de tenir compte de la corrélation transversale des moyennes de petites régions est de permettre l'existence d'une corrélation entre les résidus η_{tkj} et η_{tmj} des modèles qui décrivent la variation longitudinale des coefficients de régression β_{tkj} et β_{tmj} pour les régions k et m (équation 2.4). En règle générale, on peut raisonnablement supposer que la corrélation diminue à mesure que s'accroît la distance entre les régions. Cette relation peut être exprimée au moyen de l'équation $E(\eta_{tkj}, \eta_{tmj}) = \delta_{jj}\rho_{jj}f_j(k, m)$, où $f_j(k, m)$ est une fonction monotone décroissante des distances $D(k, m)$. La décroissance géométrique de la corrélation est exprimée par l'équation $f_j(k, m) = \rho_j^{|k-m|-1}$. Enfin, pour exprimer une corrélation fixe, nous avons la relation $f_j(k, m) \equiv 1$. C'est sur ce dernier cas que nous con-centrons notre attention maintenant. L'hypothèse d'une corrélation transversale fixe pour tous les coefficients de régression peut être traduite par l'équation

(2.6)
$$E(\eta_{tkj}\eta_{tm}') = D(\Delta)\bar{0}, \quad k \neq m$$

où $D(\Delta)$ est une matrice diagonale dont la diagonale principale est formée des variances δ_{jj} et 0 est une autre matrice diagonale formée des coefficients de corrélation ρ_j .

De même, la moyenne (de superpopulation) des valeurs de la variable étudiée pour la région k à la période t est définie

(2.2)
$$\Theta_{tk} = E(M_{tk} | \hat{g}_{tk}) = \hat{X}_{tk} \hat{g}_{tk}$$

où

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} \quad \text{et} \quad \hat{X}_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} \hat{x}'_{tki}$$

avec $i = 1, \dots, N_{tk}$ désignant les unités de la population. Evidemment, lorsque $\hat{x}'_{tki} \equiv \hat{x}_{tki}$, alors $\hat{X}_{tk} = \bar{x}'_{tk}$.

Soit \hat{g}_{tk} l'estimateur de \hat{g}_{tk} . Par conséquent, $\Theta_{tk} = \hat{X}_{tk} \hat{g}_{tk}$ et

$$M_{tk} = \frac{1}{N_{tk}} \sum_{i=1}^{N_{tk}} Y_{tki} + \sum_{i=1}^{N_{tk}} \hat{x}'_{tki} \hat{g}_{tk} \left[\Theta_{tk} + \frac{1}{N_{tk}} \left(\sum_{i=1}^{N_{tk}} (Y_{tki} - \hat{x}'_{tki} \hat{g}_{tk}) \right) \right]$$

Une caractéristique notable du modèle (2.1) est que les coefficients \hat{g}_{tk} peuvent varier de façon transversale et longitudinale. Les équations ci-dessous décrivent la variation longitudinale des coefficients:

(2.3)
$$\begin{bmatrix} \beta_{tkj} \\ \beta_{t-1,kj} \end{bmatrix} = T_j \begin{bmatrix} \beta_{tkj} \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_{tkj}, j = 0, \dots, p$$

dans ces équations, β_{tkj} , $j = 0, 1, \dots, p$, représente des coefficients fixes, que nous interprétons plus bas, et T_j représente des matrices fixes de dimension (2×2) et où les résidus $\{\eta_{tkj}\}$ satisfont les équations

(2.4)
$$E(\eta_{tkj}) = 0, E(\eta_{tkj} \eta_{t-d,kj}) = \delta_{jt}, E(\eta_{tkj} \eta_{t-d,kj}) = 0 \quad \text{pour} \quad d > 0.$$

L'expression (2.4) implique que les résidus de coefficients différents se rapportant à la même période t peuvent être corrélés mais que l'autocorrélation et l'intercorrélacion sériale sont supposées nulles.

Considérons maintenant quelques exemples d'utilisation de (2.3) assez simples:

(a) $T_j = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ implique que $\beta_{tkj} + \eta_{tkj} = \beta_{t-1,kj} + \eta_{t-1,kj}$; on reconnaît là le modèle de trajet aléatoire. (Voir, par exemple, Cooley et Prescott (1976) ainsi que LaMotte et McWhorter (1977) pour des exemples d'application de ce modèle dans des études économétriques.) Dans ce cas précis, le coefficient β_{tkj} est évidemment superflu et devrait donc être omis de sorte que $T_j \equiv 1$.

(b) $T_j = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ implique que $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$; on reconnaît là le modèle de trajet aléatoire. (Voir, par exemple, Cooley et Prescott (1976) ainsi que LaMotte et McWhorter (1977) pour des exemples d'application de ce modèle dans des études économétriques.) Dans ce cas précis, le coefficient β_{tkj} est évidemment superflu et devrait donc être omis de sorte que $T_j \equiv 1$.

(c) $T_j = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ implique que $\beta_{tkj} = \beta_{t-1,kj} + \eta_{tkj}$; on reconnaît là le modèle de trajet aléatoire. (Voir, par exemple, Cooley et Prescott (1976) ainsi que LaMotte et McWhorter (1977) pour des exemples d'application de ce modèle dans des études économétriques.) Dans ce cas précis, le coefficient β_{tkj} est évidemment superflu et devrait donc être omis de sorte que $T_j \equiv 1$.

Cependant, aucune de ces études n'a pour objet l'estimation (la prédiction) de moyennes de petites régions à partir de données d'enquête. Certains ouvrages portant sur l'estimation de moyennes de populations ont traité de l'ajustement de modèles chronologiques à des données d'enquête; voir à ce propos les articles de Binder et Dick (1989), de Tillier (1989) et de Pfeffermann (1991). Pourtant, ces méthodes ne sont pas très répandues parce que les estimateurs classiques de moyennes globales sont souvent presque aussi efficaces lorsque les modèles sont valides ou plus robustes lorsque les modèles s'avèrent non fondés.

La situation est tout à fait différente lorsque'il s'agit d'estimation pour petites régions et nous sommes d'avis que l'utilisation de modèles chronologiques peut être très avantageuse dans ce cas. Bien que la nature exacte du modèle à utiliser pour une application particulière dépende évidemment des données, la classe de modèles que nous examinons dans la section suivante est suffisamment large pour nous permettre de résoudre bon nombre, et même la plupart, des problèmes ayant trait à l'estimation pour petites régions qui surviennent en pratique. En outre, l'estimation de ces modèles a l'avantage d'être relativement simple (voir section 3).

Lorsqu'on utilise un modèle, on est toujours amené à se demander comment se prémunir contre sa défaillance et cette question devient d'autant plus délicate lorsque'il s'agit d'un modèle destiné à la production de statistiques officielles. Dans la section 4, nous nous penchons sur le problème et proposons de modifier les prédicteurs (fondés sur un modèle) de moyennes globales de petites régions de telle sorte qu'ils concordent avec les estimateurs d'enquête correspondants, pourvu que ceux-ci soient fiables. Nous examinons ensuite les propriétés statistiques des nouveaux prédicteurs. Enfin, dans la section 5, nous présentons des résultats empiriques de l'application du modèle selon deux scénarios: avec prédicteur modifié et sans prédicteur modifié. Les données utilisées dans les circonstances sont les prix de vente des maisons enregistrées à Jérusalem entre septembre 1985 et juin 1989. Le Bureau central de la statistique, en Israël, se sert régulièrement de ces données pour calculer les indices des prix du logement.

2. RÉGRESSION AVEC DES COEFFICIENTS QUI VARIENT DE FAÇON TRANSVERSALE ET LONGITUDINALE

2.1 Une classe générale de modèles

Pour les besoins de notre exposé, nous allons désigner par \tilde{Y}^k le vecteur $n^k \times 1$ des valeurs observées de la variable expliquée Y pour la région k à la période t , $k = 1, \dots, K$, $t = 1, 2, \dots$. Pour des raisons de commodité, nous supposons que $n^k \geq 1$ mais nous verrons plus loin que le modèle prévoit l'absence d'observations pour des régions à certaines périodes. Posons X_k comme la matrice de plan $n^k \times (p + 1)$ des variables auxiliaires, dont la première colonne est formée de uns. Dans de nombreuses applications, le même vecteur ligne \tilde{x}_i^k de valeurs auxiliaires s'applique à toutes les valeurs observées de Y pour une période donnée, auquel cas $X_k = \tilde{1}_{n^k} \tilde{x}_i^k$, où $\tilde{1}_{n^k}$ est un vecteur colonne formé de uns et de dimension n^k . Cela se produit lorsque les estimateurs d'enquête pour petites régions sont les seules données dont nous disposons. À cause des exigences du secret statistique et des frais de traitement, on peut difficilement utiliser les micro-données provenant d'enquêtes menées auprès de particuliers. La théorie que nous exposons ici n'est pas contrainte à la disponibilité de micro-données (voir l'exemple à la section 2.2) mais le fait qu'il existe ou non des données a un effet indéniable sur la spécification des modèles et la précision de l'estimation.

Nous définissons comme suit le modèle de régression pour la région k et la période t :

$$\tilde{Y}^k = X_k \tilde{\beta}^k + \varepsilon^k; E(\varepsilon^k) = 0, E(\varepsilon^k \varepsilon_i^k) = \sigma^2 I_{n^k} \quad (2.1)$$

ou $\tilde{\beta}^k = (\beta^{k0}, \beta^{k1}, \dots, \beta^{kp})$.

Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales

D. PFEFFERMANN et L. BURCK¹

RÉSUMÉ

Dans l'estimation pour petits domaines (ou petites régions), on cherche le plus souvent à exploiter le caractère transversal des données de manière que l'information relative à une petite région serve pour d'autres petites régions. Par ailleurs, dans le cas des enquêtes à passages répétés, il est possible d'accroître l'efficacité de l'estimation en modélisant aussi les propriétés temporelles des données. Nous expliquons notre pensée en considérant des modèles de régression à coefficients corrélés de façon transversale et qui varient de façon longitudinale. L'emploi de données de périodes antérieures pour estimer des moyennes de périodes courantes nous amène à nous interroger sur la façon de prévenir les défaillances de modèle. Dans cet article, nous proposons des modifications pour que les prédicteurs de moyennes globales de petites régions (fondés sur un modèle) concordent avec les estimateurs d'enquête correspondants et nous examinons les propriétés statistiques de ces modifications. Nous appliquons ensuite la nouvelle méthode à des données sur le prix de vente des maisons, qui servent au calcul des indices des prix du logement.

MOTS CLÉS: Filtre de Kalman; contraintes linéaires; modèles d'espace d'états.

1. INTRODUCTION

Les organismes nationaux de statistique sont souvent appelés à construire des estimateurs fiables pour les moyennes de petites régions mais la taille des échantillons à l'intérieur de ces régions est habituellement trop faible pour que l'on puisse utiliser des estimateurs directs. C'est pourquoi on a proposé ces dernières années de nouveaux estimateurs qui combinent les données d'enquête de toutes les petites régions avec de l'information supplémentaire (tirée d'un recensement ou de dossiers administratifs). Ces estimateurs ont tous pour caractéristique de pouvoir être construits, en règle générale, comme une combinaison linéaire de deux éléments: un "estimateur synthétique" de forme $\hat{X}'_i\hat{g}$, où \hat{X}_i représente l'information supplémentaire moyenne pour une petite région et \hat{g} est un vecteur de coefficients de régression estimés, et un "facteur de correction" de forme $(y_i - \hat{x}'_i\hat{g})$, où y_i et \hat{x}_i sont les moyennes observées de la variable étudiée et de la variable auxiliaire respectivement. Les facteurs de correction servent à expliquer la variation des moyennes de petites régions que n'expliquent pas les variables auxiliaires. Les divers estimateurs se distinguent principalement les uns des autres par la méthode de calcul des poids attribués aux deux éléments de la combinaison linéaire; ces méthodes vont de l'"approche fondée sur un plan" (Särndal et Hidroglou 1989) à la "méthode empirique de Bayes" (Fay et Herriot 1979), en passant par les "modèles linéaires composés" (Battese, Harter et Fuller 1989; Pfeffermann et Barnard 1991).

Il existe très peu d'ouvrages en statistique où l'on examine la possibilité d'exploiter l'aspect chronologique des données pour accroître l'efficacité des estimateurs pour petites régions et ce malgré le fait que bon nombre de ces estimateurs viennent d'enquêtes à passages répétés comme les enquêtes sur la population active. La littérature économétrique renferme un très grand nombre d'études portant sur la modélisation combinée de données chronologiques et de données transversales; voir, par exemple, Rosenberg (1973b), Johnson (1977, 1980), Maddala (1977, chapitre 7), Dielman (1983) et Pfeffermann et Smith (1985) pour des analyses.

¹ D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905. L. Burck, Unit for Statistical Analysis, Central Bureau of Statistics, Jerusalem 91130.

- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WOLTER, K.M., ISAKI, C.T., STURDEVANT, T.R., MONSOUR, N.J., et MAYES, F.M. (1976). Sample selection and estimation aspects of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 99-109.
- WOLTER, K.M., et MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (éds. D. Krewski, R. Platek et J.N.K. Rao). New York: Academic Press, 199-226.

BINDER, D.A., et DICK, J.P. (1989). Implications of survey designs for estimating seasonal ARIMA models. Article présenté à la réunion annuel de la American Statistical Association, Washington, D.C.

BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis: Forecasting Control*. San Francisco: Holden Day.

CHUNG, K.L. (1968). *A Course in Probability Theory*. New York: Harcourt, Brace World, Inc.

ELTINGE, J.L., et FULLER, W.A. (1989). Time series random component models for sample surveys. Article présenté à la réunion d'hiver de la American Statistical Association, San Diego, CA.

FULLER, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

FULLER, W.A., et ISAKI, C.T. (1981). Survey design under superpopulation models. *Current Topics in Survey Sampling*, (éds. D. Krewski, R. Platek, et J.N.K. Rao). New York: Academic Press, 199-226.

GARRETT, J.K., DETLEFSEN, R.E., et VEBUM, C.S. (1987). Recent sample revisions and related enhancements for business surveys of the U.S. Bureau of the Census. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 141-149.

GRANGER, C.W.J., et NEWBOLD, P. (1976). Forecasting transformed series. *Journal of the Royal Statistical Society, Series B*, 38, 189-203.

HAUSMAN, J.A., et WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

HILLMER, S.C., et TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.

ISAKI, C.T., WOLTER, K.M., STURDEVANT, T.R., MONSOUR, N.J., et TRAGER, M.L. (1976). Sample redesign of the Census Bureau's monthly business surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 90-98.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.

MCLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

MCLEOD, I. (1977). Correction to derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 26, 194.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.

PERFFERMAN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, à paraître.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

RAO, J.N.K., SRINATH, K.P., et QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. *Panel Surveys*, (éds. Daniel Kasprzyk, Greg Duncan, Graham Kalton et M.P. Singh). New York: Wiley, 457-479.

SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SCOTT, A.J., SMITH, T.M.F., et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.

SMITH, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. *Survey Sampling and Measurement*, (éd. N.K. Namboodiri). New York: Academic Press, 201-216.

TAM, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *Revue Internationale de Statistique*, 55, 63-73.

TRABELSI, A., et HILLMER, S.C. (1990). Benchmarking time series with reliable benchmarks. *Applied Statistics*, 39, à paraître.

d'estimation des paramètres. Deuxièmement, nous ne savons pas exactement pourquoi les estimations de l'extraction de signal se maintiennent au-dessus ou au-dessous des estimations composites pendant de longues périodes. (Cela se voit clairement dans la figure 2b, et aussi dans la figure 2a.) Dans le cas des débits de boissons, ce phénomène ressortait constamment du modèle d'échantillonnage. Nous nous attachons présentement à découvrir si ce phénomène ne serait pas imputable à une mauvaise spécification du modèle d'erreurs d'échantillonnage ou du modèle du signal. En fait, Bell et Wilcox (1990) affirment que les corrélations de e_t et e_{t-1} aux décalages qui ne sont pas des multiples de 3 ne sont pas nécessairement nulles comme le supposait le modèle.

REMERCIEMENTS

Cet article s'inspire en partie d'un travail de recherche subventionné par la National Science Foundation sous le numéro SES 84-01460 ("On-Site Research to Improve the Government-Generated Social Science Data Base") et par le U.S. Bureau of the Census et l'Université du Kansas en vertu des conventions sur la statistique n° 87-14 et 88-27. Les opinions, conclusions ou recommandations exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position de la National Science Foundation, du Census Bureau ou de l'Université du Kansas. Nous tenons à remercier Phillip Kott et David Findley pour leurs précieux commentaires concernant les résultats relatifs à la convergence, Abdelwahed Trabelsi pour le soutien qu'il nous a offert à titre d'associé de l'ASA/NSF/Census Research, de même qu'un arbitre anonyme qui, par ses commentaires, a contribué à améliorer sensiblement cet article. Nous adressons aussi des remerciements à Ruth Detlefsen, Michael Shimberg et Carol Veum, de la Business Division du Census Bureau, pour nous avoir fourni des données de la Retail Trade Survey et nous avoir renseigné sur cette enquête. Enfin, nous tenons à remercier de façon particulière James Bozik, Mark Otto et Marian Pugh, qui ont consacré beaucoup d'effort à l'élaboration du logiciel qui a servi à cette étude. Toute erreur qui pourrait s'être glissée dans l'analyse des exemples est la responsabilité des auteurs.

BIBLIOGRAPHIE

- BELL, W.R. (1984). Signal extraction for nonstationary time series. *Annals of Statistics*, 12, 646-664.
- BELL, W.R. (1987). A Note on overdifferentencing and the equivalence of seasonal time series models with monthly means and models with $(0,1,1)_{12}$ seasonal parts when $\Theta = 1$. *Journal of Business and Economic Statistics*, 5, 383-387.
- BELL, W.R., et HILLMER, S.C. (1983). Modeling time series with calendar variation. *Journal of the American Statistical Association*, 78, 526-534.
- BELL, W.R., et HILLMER, S.C. (1989). Modeling time series subject to sampling error. Research Report 89/01, Statistical Research Division, Bureau of the Census.
- BELL, W.R., et HILLMER, S.C. (1990). A Matrix approach to signal extraction for nonstationary time series models. Soumis pour publication.
- BELL, W.R., et WILCOX, D.W. (1990). The effect of sampling error on the time series behavior of consumption data. Paper presented at the CRDE/Journal of Econometrics Conference on Seasonality in Econometric Models, Montréal, Canada, mai 1990.
- BINDER, D.A., et DICK, J.P. (1986). Modeling and estimation for repeated surveys. Statistique Canada Rapport technique. Division des méthodes d'enquêtes sociales.

Tableau 5

Sensibilité des résultats par rapport à une variation de (ϕ^3, Φ) pour les débits de boissons									
(i) Valeurs de $\sigma^2_2 \times 10^5$ pour une paire de valeurs (ϕ^3, Φ) donnée									
ϕ^3									
.764	.564	.614	.664	.714	.764	Φ	.614	.664	.714
16.90	14.70	12.36	9.98	7.64	7.64		.614	.664	.714
15.03	13.00	10.87	8.72	6.62	6.62		.664	.714	.764
13.04	11.23	9.30	7.44	5.60	5.60		.714	.764	.814
10.96	9.40	7.78	6.15	4.58	4.58		.764	.814	
8.79	7.51	6.17	4.85	3.58	3.58		.814		
(ii) Valeurs de CV_{56} pour une paire de valeurs (ϕ^3, Φ) donnée									
ϕ^3									
.564	.614	.664	.714	.764	.764	Φ	.614	.664	.714
2.78	2.88	2.99	3.12	3.27	3.27		.614	.664	.714
2.95	3.04	3.14	3.26	3.38	3.38		.664	.714	.764
3.10	3.19	3.28	3.39	3.50	3.50		.714	.764	.814
3.24	3.33	3.42	3.51	3.60	3.60		.764	.814	
3.36	3.45	3.54	3.62	3.70	3.70		.814		

Nous étudions ensuite la sensibilité de CV_{56} par rapport à une variation de $Var(\log(u_i))$. Le seul paramètre du modèle d'échantillonnage qui peut être touché par une telle variation est σ^2_2 . Le tableau 4 donne les valeurs de $Var(\log(u_i))$, et de sa racine carrée ($CV(HT)$) ainsi que les valeurs correspondantes de σ^2_2 et les valeurs CV_{56} qui en découlent. CV_{56} est moins sensible aux variations que dans le tableau 3.

Enfin, nous examinons la sensibilité de CV_{56} par rapport à ϕ^3 et à Φ . Si nous posons $Var(\log(u_i))$ égale à .00776 et que nous modifions (ϕ^3, Φ), σ^2_2 variera aussi. Le tableau 5 donne les séries de valeurs utilisées pour (ϕ^3, Φ) ainsi que les valeurs de σ^2_2 et de CV_{56} qui en découlent. Notons que σ^2_2 est plus sensible ici que dans le tableau 4 et que CV_{56} augmente sensiblement lorsque ϕ^3 et Φ augmentent.

En conclusion, nous pouvons dire qu'une variation modérée des paramètres du modèle d'échantillonnage influe relativement peu sur θ_i – les plus fortes variations observées chez θ_i étant de l'ordre de 2% – mais a des effets plus notables sur les variances de l'extraction de signal, la variation des valeurs de CV_{56} pouvant aller jusqu'à 17%. Cela porte à croire qu'en ce qui concerne cet exemple, la chose dont il faut se préoccuper le plus lorsqu'on ne connaît pas les paramètres du modèle d'erreurs d'échantillonnage est probablement l'effet que cela peut avoir sur la variance de l'extraction de signal de même que les mesures permettant d'évaluer la différence de qualité par rapport aux estimations composites. Néanmoins, pour tous les cas étudiés dans l'analyse de sensibilité, nous avons observé une amélioration notable de la variance dans le cas des estimations d'extraction de signal.

5.4 Conclusions

L'exemple des débits de boissons illustre bien les avantages que l'on peut tirer de l'application de l'approche chronologique à l'estimation dans les enquêtes à passages répétés. Les deux exemples donnent aussi un aperçu du caractère délicat et complexe de la modélisation de séries chronologiques que l'on peut être appelé à faire. Nous considérons ces résultats comme provisoires pour plusieurs raisons. Premièrement, nous avons déjà souligné le caractère "optimiste" des variances d'extraction de signal, qui ne tiennent pas compte de l'erreur

En comparant les modèles (5.8b) et (5.9b), nous pouvons avoir une idée de la sensibilité du modèle du signal par rapport à une variation de σ^2_{ϵ} , la variance de résidus du modèle d'échantillonnage, puisque (5.8b) correspond à $\sigma^2_{\epsilon} = 0$ et (5.9b), à $\sigma^2_{\epsilon} = 9.3 \times 10^{-5}$. Nous observons les variations les plus notables dans la valeur estimée de σ^2_{ϵ} , ce qui n'est pas étonnant, et dans la valeur estimée du paramètre de moyenne mobile saisonnière, η_{12} par exemple, qui était essentiellement égal à 1 lorsqu'on a formulé le modèle (5.9b). La réestimation du modèle du signal pour d'autres valeurs de σ^2_{ϵ} a donné $\eta_{12} \geq .99$ pour $\sigma^2_{\epsilon} \geq 3.0 \times 10^{-5}$. Compte tenu de cela, et pour simplifier la présentation des résultats, nous supposons que $\eta_{12} = 1$ et utilisons un modèle du signal qui, comme (5.9b), comprend des variables indicatrices saisonnières. La figure 2c donne les estimations d'extraction de signal θ_1 (désaisonnalisées et corrigées en fonction de la variation des jours commerciaux) qui correspondent aux modèles d'erreurs d'échantillonnage pour lesquels $(\phi_3, \Phi) = (.564, .614)$ et $(.764, .814)$ et où $\rho = .986$ et $\text{Var}(\log(u_t)) = .00776$ (la variance relative des estimations d'Horvitz-Thompson) sont fixes. Ces estimations comprennent les valeurs extrêmes de θ_1 pour l'analyse de sensibilité. La nature des diverses estimations θ_1 que nous avons produites semble correspondre à peu près à la valeur de $\text{CV}_{56} = [\text{Var}(\log(\theta_{56}))]^{1/2}$, qui est le coefficient de variation de l'extraction de signal observé au milieu de la série. (CV_{56} est très proche de la valeur minimum, qui est observée à $t = 53$ - voir tableau 2.) Moins CV_{56} est élevé, plus θ_1 est lisse. CV_{56} est 2.78% et 3.28% et 3.70% pour des valeurs (ϕ_3, Φ) de (.564, .614), (.664, .714) et (.764, .814) respectivement. D'autres estimations θ_1 que nous avons produites sont encore plus près des estimations de l'extraction de signal des figures 2b ou 2c, l'écart le plus mince étant observé à CV_{56} . Nous allons maintenant étudier la sensibilité de CV_{56} par rapport à une variation des paramètres du modèle d'erreurs d'échantillonnage, à commencer par ρ . Le seul paramètre de (5.7b) qui soit touché par une variation de ρ est η . Le tableau 3 donne les valeurs de η considérées et les valeurs de ρ correspondantes ainsi que les valeurs CV_{56} qui en découlent. Nous constatons que CV_{56} est assez sensible aux variations de ρ , plus particulièrement aux hausses: CV_{56} est 6% plus élevé pour $\rho = 1$ (3.49) que pour $\rho = .985$ (3.28), qui est la valeur utilisée pour (5.7b).

Tableau 3

Sensibilité de CV_{56}^1 par rapport à une variation de η (variation de ρ) pour les débits de boissons

η	ρ	CV_{56}
.00	.9375	3.03
-.05	.9642	3.12
-.10	.9792	3.21
-.15	.9888	3.31
-.20	.9953	3.40
-.25	1.000	3.49

¹ CV_{56} est le coefficient de variation de l'extraction de signal pour $t = 56$ (milieu de la série), exprimé en pourcentage, c.-à.-d. racine carrée de $\text{Var}(\log(\theta_1)) - \log(\theta_1)$ multipliée par 100.

Tableau 4

Sensibilité de CV_{56} par rapport à une variation de $\text{Var}(\log(u_t))$ ¹ (variation de σ^2_{ϵ}) pour les débits de boissons

$\text{Var}(\log(u_t))$	$\text{CV}(\text{HT})^2$	$\sigma^2_{\epsilon} \times 10^5$	CV_{56}
.00676	8.22	8.16	3.16
.00726	8.52	8.76	3.23
.00776	8.81	9.30	3.28
.00826	9.09	9.97	3.35
.00876	9.36	10.57	3.40

¹ $\text{Var}(\log(u_t))$ est la variance relative des estimateurs d'Horvitz-Thompson.

² $\text{CV}(\text{HT})$ est le coefficient de variation des estimateurs d'Horvitz-Thompson, exprimé en pourcentage, c.-à.-d. racine carrée de $\text{Var}(\log(u_t))$ multipliée par 100.

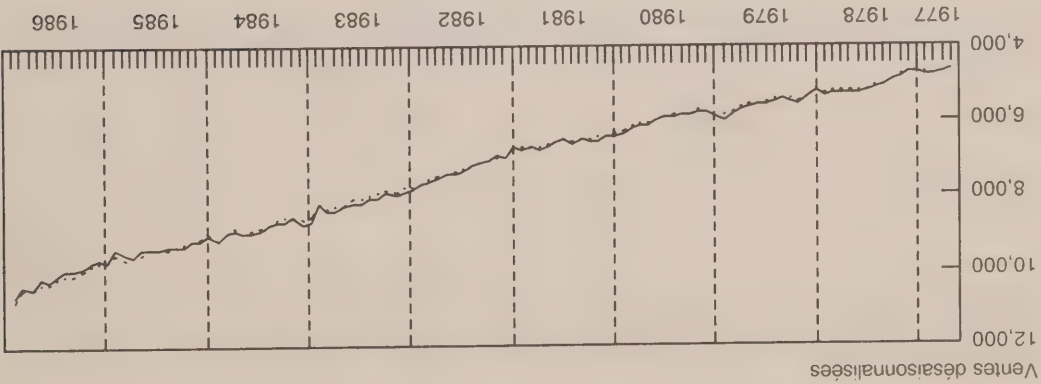


Figure 2.a Restaurants: estimations composites (continu) et estimations de l'extraction de signal (pointillé)

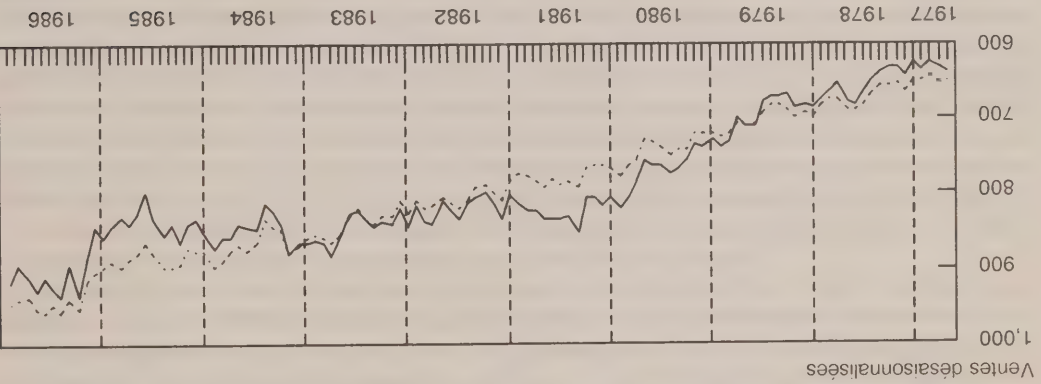


Figure 2.b Débits de boissons: estimations composites (continu) et estimations de l'extraction de signal (pointillé)

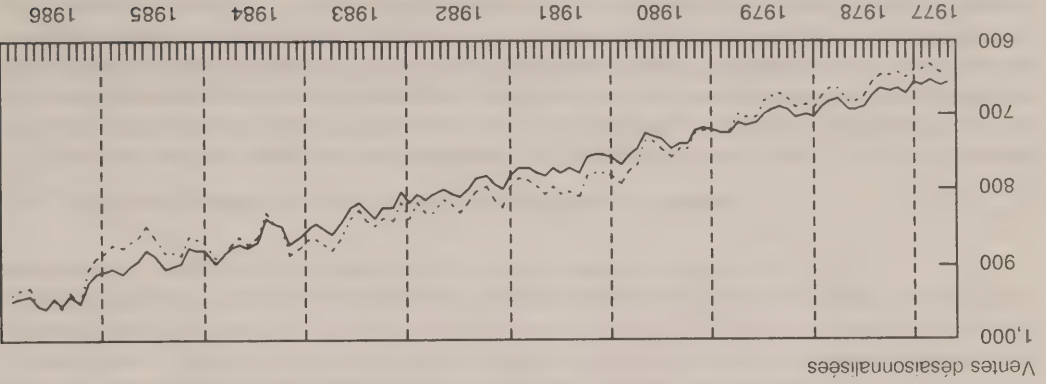


Figure 2.c Débits de boissons: autre série d'estimations de l'extraction de signal

ainsi qu'une fonction de régression saisonnière de la forme $\sum_{i=1}^I \gamma_i M_{it}$, où M_{it} est 1 en janvier, -1 en décembre et 0 aux autres mois, \dots , M_{11t} est 1 en novembre, -1 en décembre et 0 aux autres mois, \dots , M_{11t} est 1 en novembre, -1 en décembre et 0 aux autres mois. On obtient ainsi les modèles estimés suivants:

$$(1-B) \left[\log(\theta_t) - \sum_{i=1}^I \beta_i T_{it} - \sum_{i=1}^I \gamma_i M_{it} \right] = .00762 + (1 - .20B - .29B^2)b_t$$

(restaurants)

$$\hat{\sigma}_b^2 = .000139$$

(5.9a)

$$(1-B) \left[\log(\theta_t) - \sum_{i=1}^I \beta_i T_{it} - \sum_{i=1}^I \gamma_i M_{it} \right] = .00352 + (1 - .18B - .09B^2 - .42B^3)b_t$$

(débits de boissons)

$$\hat{\sigma}_b^2 = .000244$$

(5.9b)

Là encore, nous omettons les valeurs estimées des paramètres de régression. Nous ne calculons pas non plus d'erreurs types pour les paramètres ARMA; compte tenu des modèles que nous utilisons ici, ce genre de calcul mérite de faire l'objet de recherches. Toutefois, l'hypothèse peu réaliste selon laquelle le modèle d'échantillonnage est connu rend la tâche particulièrement difficile dans notre contexte. En examinant les résidus normalisés obtenus au moyen du filtre de Kalman, ainsi que les autocorrélations correspondantes, on ne trouve aucune évidence majeure dans les modèles ajustés pour l'une ou l'autre série.

Nous nous sommes servis des modèles estimés (5.7a,b) et (5.9a,b) pour établir des estimations d'extraction de signal de $\log(\theta_t)$, que nous avons ensuite retransformées pour obtenir les valeurs estimées de θ_t . Les résultats pertinents pour les deux séries sont reproduits dans les figures 2a et 2b, abstraction faite des effets saisonniers estimés et des effets de la variation des jours commerciaux. Notons que l'extraction de signal produit des estimations très peu différentes des estimations composées pour ce qui est des restaurants, l'erreur d'échantillonnage étant d'ailleurs assez faible dans leur cas (variance relative faible); en revanche, la même extraction produit des estimations très différentes des estimations composées dans le cas des débits de boissons, pour lesquels l'erreur d'échantillonnage est beaucoup plus forte (variance relative plus élevée). Nous avons aussi calculé des variances d'extraction de signal pour $\log(\theta_t)$; il s'agit des variances relatives des valeurs estimées de θ_t . Le tableau 2 montre que, dans le cas des restaurants, l'extraction de signal produit des CV de 8 à 32% moins élevés – selon le mois auquel correspond l'estimation dans la série – que ceux des estimateurs composites finals (qui sont déjà peu élevés) et que dans le cas des débits de boissons, elle produit des CV de 27 à 38% moins élevés. Comme nous l'avons déjà mentionné, ces résultats sont optimistes car ils supposent que les véritables modèles de composantes sont ceux qui ont été estimés. Afin de voir en partie de quoi il retourne, nous allons maintenant étudier la sensibilité des résultats pour les débits de boissons par rapport à une variation des paramètres des modèles.

5.3 Analyse de sensibilité pour les ventes des débits de boissons

Dans cette section, nous nous penchons sur la sensibilité des résultats par rapport à des variations dans le modèle d'échantillonnage, étant donné que celui-ci a été défini avec moins de renseignements que le modèle du signal. Notre méthode consiste à faire varier des paramètres du modèle d'erreurs d'échantillonnage, puis à réestimer le modèle du signal et à refaire l'extraction de signal. Bien qu'il serait préférable de disposer de mesures statistiques plus formelles pour rendre compte de l'erreur d'extraction de signal attribuable à l'erreur de modèle (ce qui n'est pas possible dans l'état actuel de la théorie et des logiciels), cette méthode devrait à tout le moins indiquer sous quels rapports les résultats de l'extraction de signal sont sensibles à une variation de paramètres et sous quels rapports ils ne le sont pas.

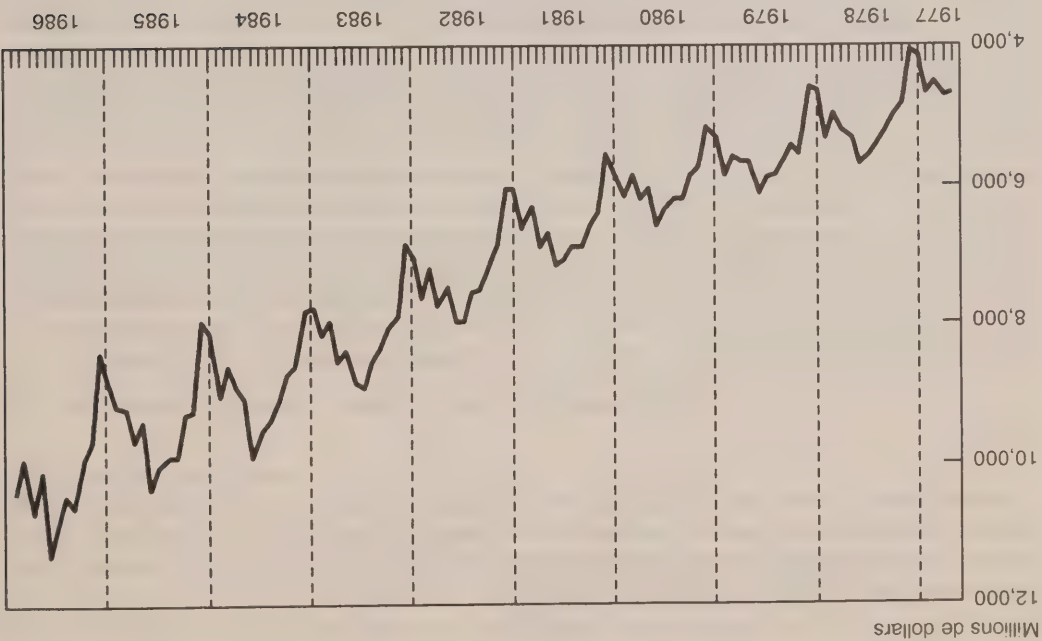


Figure 1.a Ventes au détail des restaurants — estimations composites (non étalonnées)

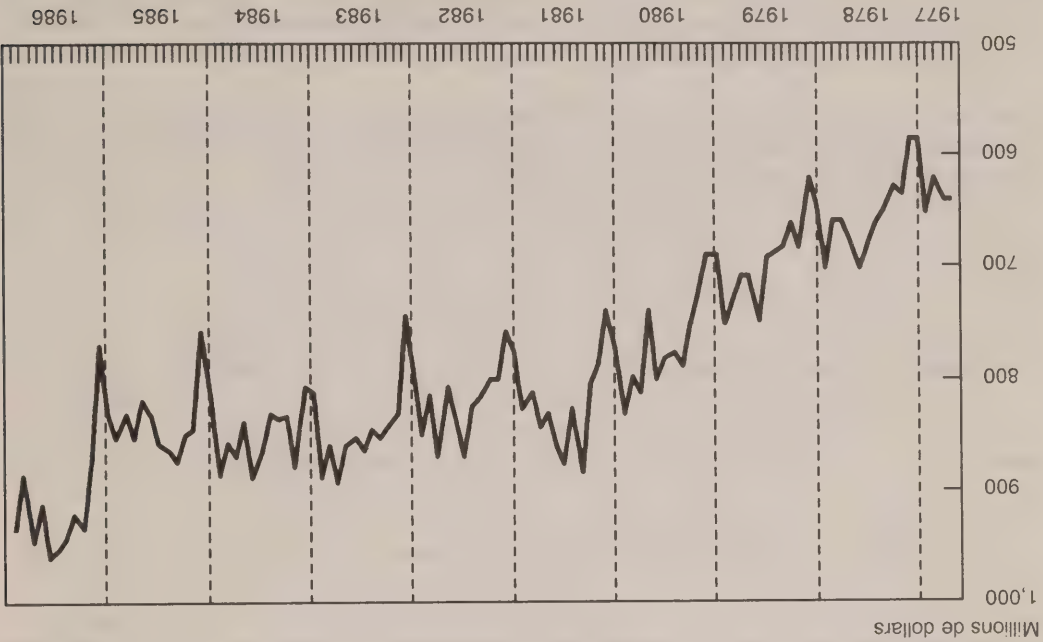


Figure 1.b Ventes au détail des débits de boissons — estimations composites (non étalonnées)

On peut employer la méthode de McLeod (1975, 1977) pour déterminer $\text{Var}(\log(u_i))$ dans ces modèles; $\text{Var}(\log(u_i))$ est une estimation de la variance relative de l'estimateur composite final. Le tableau 2 donne les coefficients de variation correspondants (.025 pour les restaurants et .052 pour les débits de boissons). Ces valeurs sont très proches des estimations qui figurent dans les rapports mensuels sur le commerce de détail du Census Bureau et qui sont obtenues d'une manière plus directe.

5.2 Modélisation de séries chronologiques et extraction de signal

Les figures 1a et 1b montrent le graphique de la série des estimations composites finales Y_t pour les restaurants et les débits de boissons respectivement. Afin d'élaborer des modèles pour les restaurants et les débits de boissons, nous avons pensé que l'opérateur de différence empiriques pour $\log(Y_t)$ et ses différences, nous avons pensé que l'opérateur de différence $(1 - B)(1 - B^{12})$ serait approprié pour les deux séries. On sait que les séries touchant le commerce de détail sont influencées par la variation des jours commerciaux; il est possible de modéliser cette variation en introduisant dans le modèle sept variables de régression: $X_{1t} =$ nombre de lundis dans le mois t , \dots , $X_{7t} =$ nombre de dimanches dans le mois t . D'après Bell et Hillmer (1983), il serait plus commode d'utiliser les variables $T_{1t} = X_{1t} - X_{7t}$ (nombre de lundis - nombre de dimanches), \dots , $T_{6t} = X_{6t} - X_{7t}$ (nombre de samedis - nombre de dimanches), $T_{7t} = \sum_{i=1}^7 X_{it}$ (longueur du mois t). Afin de définir les structures ARMA, nous avons analysé les autocorrélations et les autocorrélations partielles des résidus de la régression de $(1 - B)(1 - B^{12}) \log(Y_t)$ par rapport à $(1 - B)(1 - B^{12})T_{it}$, $i = 1, \dots, 7$. Cette analyse nous a amenés à définir un modèle ARMMI(0,1,2)(0,1,1)¹² pour les restaurants et un modèle ARMMI(0,1,3)(0,1,1)¹² pour les débits de boissons. Nous avons donc les modèles suivants:

$$(1 - B)(1 - B^{12}) \left[\log(Y_t) - \sum_{i=1}^7 \beta_i T_{it} \right] = (1 - .25B - .22B^2)(1 - .79B^{12}) a_t \quad (5.8a)$$

(restaurants) $\hat{\sigma}_a^2 = .000230$

$$(1 - B)(1 - B^{12}) \left[\log(Y_t) - \sum_{i=1}^7 \beta_i T_{it} \right] = (1 - .21B - .15B^2 + .03B^3)(1 - .56B^{12}) a_t \quad (5.8b)$$

(débits de boissons) $\hat{\sigma}_a^2 = .000587$

Pour des raisons de concision, nous allons omettre les valeurs estimées des paramètres de la variation des jours commerciaux. Bien que les paramètres de moyenne mobile pour les décalages 2 et 3 dans l'équation (5.8b) soient peu élevés, nous allons les conserver car les équations (5.8a) et (5.8b) ne vont servir qu'à la première étape de la modélisation de $\log(\theta_t)$ pour les deux séries.

En prenant des modèles du type (5.8a) et (5.8b) pour $\log(\theta_t)$ et les modèles (5.7a) et (5.7b) pour $\log(u_t)$, nous avons pu estimer les paramètres des modèles pour $\log(\theta_t)$. La valeur estimée des paramètres de moyenne mobile saisonnière est très proche de 1 pour les deux séries (.985 pour les restaurants et .992 pour les débits de boissons), ce qui suppose un mouvement saisonnier quasi déterministe, que l'on peut modéliser en supprimant le facteur $(1 - B^{12})$ de part et d'autre de l'équation du modèle de θ_t et en lui substituant une constante de tendance

Tableau 2

Coefficients de variation (CV)¹ pour les estimations de ventes au détail

Extraction de signal ³	Horvitz-Thompson		Composite final ²		max	
	CV		CV		min	
Restaurants	.042		.025		.017	
Débîts de boissons	.088		.052		.032	
					.038	

¹ CV = (variance relative)^{.5}.

² Les valeurs pour l'estimateur composite final sont déterminées à l'aide des modèles (5.7a) et (5.7b).

³ En fait, les valeurs pour l'extraction de signal varient dans le temps, la valeur la plus élevée se trouvant à la fin de la série et la valeur la plus faible, vers le milieu. Nous donnons ici la valeur la plus élevée (décembre 1986) et la plus faible (janvier 1982) pour les deux séries. Les variances de l'extraction de signal ne sont pas chronologiquement symétriques parce que les "rééchantillonnages" de janvier 1982 ne coïncide pas parfaitement avec le milieu de la série.

Le membre de droite de l'équation ci-dessus représente un processus de moyennes mobiles du premier ordre (Box et Jenkins, 1976, p. 121), dont on peut déterminer les paramètres en connaissant les valeurs estimées de σ_e^2 et ρ . L'équation (5.5) donnerait donc un modèle ARMA pour e_t .

Au lieu de poursuivre dans la même voie, nous allons maintenant poser l'hypothèse plutôt audacieuse qu'il existe un modèle du même type pour $\log(u_t)$ dans l'expression $\log(Y'_t) = \log(\theta_t) + \log(u_t)$, par conséquent,

(5.6) $(1 - .75B)(1 - \phi^3B^3)(1 - \Phi B^{12}) \log(u_t) = (1 - \eta B) c_t.$

Nous faisons cela parce que les estimations de la variance d'échantillonnage pour ces séries dépendent fortement du niveau de la série; les estimations de la variance relative, elles, sont beaucoup plus stables. Nous supposons de plus que les valeurs estimées de la variance relative servir à calculer η et σ_e^2 . Les valeurs estimées Y'_t , Y'_{t-1} , $\text{Var}(e'_t)$ et $\text{Var}(e'_{t-1})$ sont connues pour les années 1982 à 1986 inclusivement. Une fois les estimations de la variance relative calculées, elles ont servi à un processus d'estimation s'inspirant de la méthode du maximum de vraisemblance appliquée à la distribution normale logarithmique - c.-à-d. calculer la moyenne des logarithmes des estimations de la variance relative, additionner cette moyenne et la moitié de la variance empirique des estimations de forme logarithmique, puis prendre la valeur exponentielle de cette somme. (On aurait obtenu des résultats semblables en faisant simplement la moyenne des estimations de la variance relative.) $\text{Var rel}(Y'_t)$ et $\text{Var rel}(Y'_{t-1})$ ont été calculées de façon indépendante; on a ensuite fait la moyenne des deux pour obtenir une estimation unique de la variance relative qui est constante. Les résultats pertinents figurant dans la première colonne du tableau 2. Nous pouvons maintenant nous servir de ces résultats et des valeurs $\hat{\rho}$ calculées précédemment pour déterminer η et σ_e^2 dans le membre de droite de l'équation (5.6). Les modèles d'échantillonnage sont donc

(5.7a) $(1 - .75B)(1 - .685B^3)(1 - .723B^{12}) \log(u_t) = (1 + .130B) c_t$

(restaurants) $\hat{\sigma}_e^2 = 1.948 \times 10^{-5}$

(5.7b) $(1 - .75B)(1 - .664B^3)(1 - .714B^{12}) \log(u_t) = (1 + .134B) c_t$

(débîts de boissons) $\hat{\sigma}_e^2 = 9.301 \times 10^{-5}$

où $m = 4$ pour l'enquête à quatre panels. Ce modèle vaut aussi pour e'_{i-1} , sauf que $v_{2,i-1}$ est substituée à v_{1i} . (v_{1i} et $v_{2,i-1}$ représentent le bruit blanc avec comme variance σ_v^2 .)

Une caractéristique particulièrement utile du modèle (5.1) est que s'il peut décrire, mois après mois, l'erreur d'échantillonnage dans chaque panel avec $m = 1$, alors pour n'importe quel nombre m (qui est un diviseur de 12) de panels indépendants qui participent successivement à l'enquête, e'_i est conforme au modèle (5.1). Grâce à cette caractéristique, nous pouvons nous servir des résultats du tableau 1 pour l'enquête à quatre panels afin d'estimer ϕ^4 et Φ (en supposant que $\phi > 0$) de transformer ces valeurs en valeurs estimées de ϕ^3 et Φ , qui sont les paramètres du modèle pour l'enquête à trois panels. Nous avons donc déterminé ϕ^4 et Φ pour minimiser la somme des carrés des écarts entre les corrélations calculées à l'aide de (5.1) et celles calculées à l'aide de la transformation de Fisher. (Nous avons exclu les décalages 20 et 24 du calcul et attribué un poids de 0,5 au décalage 16 à cause du faible nombre de corrélations estimées ayant servi au calcul d'une moyenne pour ces décalages.) Nous avons ainsi obtenu les valeurs estimées $\phi^3 = .685$ et $\Phi = .723$ pour les restaurants et $\phi^3 = .664$ et $\Phi = .714$ pour les débits de boissons. Les corrélations calculées à l'aide de (5.1) pour $m = 4$ figurent dans le tableau 1 et peuvent être comparées à celles qui ont servi au calcul de moyennes. Nous pourrions envisager des méthodes d'estimation statistique plus formelles pour ϕ^3 et Φ et imaginer un test de validité de l'ajustement du modèle s'il était possible d'établir des estimations d'autocorrélation des erreurs d'échantillonnage à l'aide de micro-données plus récentes provenant de l'enquête à trois panels.

Nous posons aussi comme hypothèse que $\text{Corr}(e'_i, e'_{i-1-k}) = \rho \text{Corr}(e'_i, e'_{i-k})$ pour tous k . Pour appuyer cette hypothèse, signalons que la régression de e'_{i-1-k} par rapport à e'_{i-k} pour la population s'exprime par $\rho e'_{i-k} + \epsilon_i$; si ϵ_i et e'_i ne sont pas non corrélées, ϵ est, à tout le moins, sûrement faible puisque $\text{Var}(\epsilon) = (1 - \rho^2)\text{Var}(e'_i)$ et que ρ est très voisin de 1. Compte tenu de cette hypothèse, nous pouvons définir le modèle bidimensionnel suivant pour

(e'_i, e'_{i-1}) à partir de l'équation (5.1):

$$(1 - \phi^3 B^3)(1 - \Phi B^{12}) \begin{bmatrix} e'_i \\ e'_{i-1} \end{bmatrix} = \begin{bmatrix} v_{1i} \\ v_{2,i-1} \end{bmatrix} \quad \text{Var} \begin{bmatrix} v_{1i} \\ v_{2,i-1} \end{bmatrix} = \sigma_v^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad (5.2)$$

où $\rho = \text{Corr}(v_{1i}, v_{2,i-1}) = \text{Corr}(e'_i, e'_{i-1})$. On produit régulièrement des estimations de $\text{Corr}(e'_i, e'_{i-1})$; d'ailleurs, nous disposons de ces estimations pour les années 1982 à 1986. En faisant la moyenne de ces valeurs estimées (à l'aide de la transformation de Fisher), nous avons obtenu $\hat{\rho} = .985$ pour les restaurants et $\hat{\rho} = .986$ pour les débits de boissons.

Nous pouvons maintenant nous servir de (5.2) pour élaborer un modèle d'erreur d'échantillonnage pour l'estimateur composite de forme linéaire (Wolter 1979), qui est défini

$$Y'_{i-1} = (1 - \beta)Y'_i + \beta(Y'_{i-1} + Y'_i - Y'_{i-1}) \quad (\text{estimateur provisoire}),$$

$$Y'_{i-1} = (1 - \alpha)Y'_{i-1} + \alpha Y'_{i-1} \quad (\text{estimateur final}). \quad (5.3)$$

Dans l'enquête à trois panels, $\alpha = .8$ et $\beta = .75$. On voit facilement que les équations ci-dessus s'appliquent aussi bien aux erreurs d'échantillonnage (e étant substitué à Y). Ceci nous permet d'exprimer e'_i en fonction de e'_{i-1} et de e'_{i-1} par l'équation suivante:

$$(1 - .75B)e'_i = .2e'_{i-1} - .75e'_{i-1} + .8e'_i. \quad (5.4)$$

En combinant (5.2) et (5.4), nous obtenons

$$(1 - .75B)(1 - \phi^3 B^3)(1 - \Phi B^{12})e'_i = .2v_{2i} - .75v_{2,i-1} + .8v_{1i}. \quad (5.5)$$

Tableau 1
Corrélations d'erreurs d'échantillonnage pour les estimations d'Horvitz-Thompson

Décalage						
		4	8	12	16	20
		4	8	12	16	20
Restaurants	D'après le calcul d'une moyenne ¹	.72	.71	.79	.63	.65
	D'après (5.1) ²	.75	.69	.81	.60	.53
Débits de boissons	D'après le calcul d'une moyenne ¹	.70	.67	.78	.60	.60
	D'après (5.1) ²	.72	.66	.80	.56	.50
Nombre de corrélations ayant servi au calcul de moyennes		23	19	15	11	7
Poids utilisés dans le calcul de ϕ		1	1	1	.5	0

¹ Il existait des estimations brutes de $\text{Corr}(e'_t, e'_j)$ et de $\text{Corr}(e'_{t-1}, e'_{j-1})$ pour toutes les paires de mois de janvier 1973 à mars 1975. On a fait la moyenne des corrélations pour les décalages indiqués après avoir appliqué la transformation de Fisher puis transformé à nouveau les résultats ainsi obtenus.

² Corrélations calculées à l'aide du modèle (5.1) pour $m = 4$ avec comme paramètres $\phi^4 = .604$, $\phi^{12} = .723$ (restaurants) et $\phi^4 = .580$, $\phi^{12} = .714$ (débits de boissons). Ces valeurs ont été déterminées à l'aide des poids figurant dans le tableau pour minimiser la somme pondérée des carrés des écarts entre les corrélations calculées à l'aide de (5.1) et celles ayant servi au calcul de moyennes. Les décalages 20 et 24 n'ont pas servi au calcul (poids nul) à cause du faible nombre de corrélations estimées correspondant à ces décalages.

5.1 Elaboration de modèles d'erreur d'échantillonnage

Nous allons, en premier lieu, élaborer un modèle pour la structure de corrélation des erreurs d'échantillonnage. Exprimons par $X'_t = \theta'_t + e'_t$ l'estimation HT pour le mois courant (t) et par $X'_{t-1} = \theta'_{t-1} + e'_{t-1}$ l'estimation HT pour le mois précédent ($t - 1$). Nous allons utiliser le même modèle pour e'_t et e'_{t-1} . Les valeurs estimées de $\text{Corr}(e'_t, e'_{t-1})$ sont extrêmement élevées – en règle générale, 0.98 et plus. Bien que cette forte corrélation soit en partie artificielle (à cause des entreprises qui déclarent le même chiffre pour les ventes du mois courant et celles du mois précédent et à cause, peut-être, de la façon dont on impute les valeurs manquantes), il est difficile de faire la distinction entre les caractéristiques de e'_t et celles de e'_{t-1} lorsqu'on ne dispose pas d'autres renseignements.

Comme les trois panels avec renouvellement sont tirés de façon à peu près indépendante (Wolter 1979), les autocorrélations et les corrélations croisées pour (e'_t, e'_{t-1}) devraient être non nulles uniquement pour les décalages qui sont des multiples de 3. On peut faire la moyenne des valeurs estimées de ces corrélations avec décalage pour plusieurs périodes successives en supposant que la série est stationnaire en corrélation. Bien qu'en règle générale, on ne produise pas de valeurs estimées de corrélations avec décalage pour la Retail Trade Survey, une étude spéciale a permis de le faire au moyen de micro-données (totaux de groupes aléatoires) de l'échantillon de l'enquête pour la période allant de janvier 1973 à mars 1975. À cette époque toutefois, le plan de sondage prévoyait quatre panels avec renouvellement à l'exclusion du panel aréolaire). Faute de données plus récentes, nous avons "fait la moyenne" des corrélations aux décalages 4, 8, 12, 16, 20 et 24 pour e'_t et e'_{t-1} . (En fait, nous avons commencé par appliquer la transformation de Fisher, $.5 \log((1 + r)/(1 - r))$, pour rendre la distribution des corrélations transformées plus symétrique, puis nous avons transformé à nouveau les résultats ainsi obtenus.) Les résultats pertinents figurent dans le tableau 1. Ce tableau révèle une corrélation positive assez forte entre les erreurs d'échantillonnage et l'existence d'un mouvement saisonnier au décalage 12. Compte tenu de ces données, nous pouvons proposer le modèle suivant:

$$(1 - \phi_m B_m)(1 - \Phi B_{12})e'_t = v_{1t}, \quad (5.1)$$

définitif peut ensuite servir à l'estimation d'extraction de signal de θ . Les calculs relatifs à l'ajustement de modèle et à l'extraction de signal sont complexes; Bell et Hillmer (1989) font une analyse des algorithmes du filtre de Kalman. Ces algorithmes ont trouvé une application dans un logiciel élaboré récemment avec le concours de spécialistes de l'analyse chronologique de la Statistical Research Division du U.S. Bureau of the Census. C'est ce logiciel dont nous nous servons dans la section suivante.

5. EXEMPLE: U.S. RETAIL TRADE SURVEY - VENTES DES RESTAURANTS ET DES DÉBITS DE BOISSONS

À titre d'exemple, nous allons analyser la série chronologique des ventes (en millions de dollars) des établissements de restauration (ou restaurants) et des débits de boissons, qui sont estimées dans l'enquête mensuelle sur le commerce de détail aux E.-U. L'échantillon de la Retail Trade Survey est composé d'un panel de grandes entreprises qui sont prélevées dans un répertoire avec une probabilité égale à un et qui déclarent leurs ventes mensuellement, et de trois panels avec renouvellement formés d'entreprises de moindre importance qui sont prélevées dans un répertoire suivant un échantillonnage aléatoire simple stratifié. Notons aussi l'existence d'un panel aréolaire avec renouvellement, formé d'entreprises qui ne figurent pas à un répertoire. À chaque trimestre, on introduit dans l'échantillon global un certain nombre d'entreprises nouvelles créées, puis on supprime du même échantillon les entreprises réputées disparues selon les derniers rapports d'activité. Les entreprises qui font partie des panels avec renouvellement déclarent leurs ventes du mois courant et du mois précédent, les unes (panels de répertoire) à tous les 3 mois, les autres (panel aréolaire) à tous les 6 ou 12 mois. On établit alors des estimations d'Horvitz-Thompson (HT) des ventes du mois courant et du mois précédent; les séries chronologiques qui en découlent sont désignées par Y'_t et Y'_{t-1} . Ces séries servent ensuite à construire des estimateurs composites de la manière décrite dans Wolter (1979). Enfin, les estimations composites ainsi obtenues forment la série chronologique X_t . (Bien qu'il serait intéressant d'analyser directement Y'_t et Y'_{t-1} au lieu de faire les opérations ci-dessus, les estimations qui forment ces séries ne sont pas conservées suffisamment longtemps pour pouvoir servir à la modélisation de séries chronologiques saisonnières.) Les variances d'échantillon-nage sont estimées au moyen de la méthode des groupes aléatoires (Wolter 1985, chapitre 2), dans le cas des panels de répertoire (16 groupes), ou de la méthode de regroupement des strates, dans le cas du panel aréolaire. Pour plus de détails sur la Retail Trade Survey, voir Isaki et coll. (1976); Wolter et coll. (1976); Wolter (1979); Garrett, Delftsen et Veum (1987) ainsi que Bell et Wilcox (1990).

La Retail Trade Survey présente plusieurs contraintes pour notre analyse. Premièrement, le plan de sondage est remanié et l'échantillon reconstitué à tous les cinq ans environ, les dernières révisions ayant eu lieu en septembre 1977, en janvier 1982 et en janvier 1987. Cela a pour effet de créer un bris de continuité dans la structure de covariances de e_t à tous les cinq ans; le filtre de Kalman peut remédier à cette situation, comme l'indiquent Bell et Hillmer (1989). Nous utiliserons des données pour la période allant de septembre 1977 à décembre 1986, de sorte qu'une reconstitution de l'échantillon coïncide à peu près avec le milieu de notre série. Deuxièmement, lorsqu'un nouvel échantillon d'entreprises est introduit, on se sert d'ELNVM approximatifs pour les trois premiers mois, avant d'utiliser les estimateurs composites (Wolter 1979). Cela crée un effet transitoire dans les autocorrélations d'erreurs d'échantillon-nage, effet dont nous ne tiendrons pas compte. Enfin, les estimations mensuelles de l'enquête sont ajustées en fonction de totaux annuels établis à la suite d'une enquête annuelle et du recensement économique quinquennal. Pour éliminer cette contrainte, nous allons nous servir de données qui ne sont pas étalonnées. Toutefois, il ne faudra pas s'étonner que les données que nous utilisons ici ne concordent pas avec les estimations publiées.

moyenne des autocorrélations estimées de e_t par rapport à t pour estimer $\rho''(k)$, qui équivaut à peu près à l'autocorrélation de $\log(u_t)$. On peut, comme avant, faire la moyenne des estimations de la variance relative.

En réalité, les valeurs estimées de variances et d'autocovariances seront le plus souvent des estimations de $\text{Var}(e_t | \hat{Q})$ et de $\text{Cov}(e_t, e_{t+k} | \hat{Q})$. Elles pourront peut-être servir d'estimations pour $\text{Var}(e_t)$ et $\text{Cov}(e_t, e_{t+k})$ si, par exemple, elles peuvent convenir dans une perspective de modèle. Dans le cas contraire, et si Y_t est non biaisé selon le plan de sorte que $E(e_t | \hat{Q}) = 0$, il ne sera pas moins justifié de faire la moyenne des estimations d'autocovariances par rapport à t . Premièrement, si nous supposons que e_t est stationnaire, alors $\gamma_e(k) \equiv \text{Cov}(e_t, e_{t+k}) = E[\text{Cov}(e_t, e_{t+k} | \hat{Q})]$; nous pouvons donc calculer la moyenne des valeurs estimées de $\text{Cov}(e_t, e_{t+k} | \hat{Q})$ pour estimer $\gamma_e(k)$. Ou encore, si e_t est stationnaire en covariance relative, $E(u_t | \hat{Q}) = E(e_t | \hat{Q})/\theta_t = 0$; par conséquent, nous avons $\gamma''(k) \equiv \text{Cov}(u_t, u_{t+k}) = E[\text{Cov}(u_t, u_{t+k} | \hat{Q})] = \text{Cov}(\log(u_t), \log(u_{t+k})) + O_p(r_t^2)$, et nous pouvons faire la moyenne des valeurs estimées de $\text{Cov}(u_t, u_{t+k} | \hat{Q})$ pour estimer $\gamma''(k)$. Quant au calcul de la moyenne des valeurs estimées des corrélations conditionnelles (par rapport à \hat{Q}), il est plus difficile d'en démontrer le bien-fondé puisque $E[\text{Corr}(e_t, e_{t+k} | \hat{Q})] \neq \text{Corr}(e_t, e_{t+k})$; cependant, l'inverse pourrait être vrai si nous posons certaines hypothèses. En général, les méthodes d'estimation des structures d'autocovariances des erreurs d'échantillonnage nécessitent une étude plus poussée.

À partir d'une structure estimée des covariances des erreurs d'échantillonnage et de renseignements pertinents sur le plan de l'enquête, nous pouvons tenter de définir un modèle chronologique (y compris les paramètres) qui reproduira fidèlement cette structure. C'est ce à quoi nous nous attachons dans la section 5.

Nous allons maintenant élaborer un modèle pour le signal, θ_t . Comme le signal est ce qui explique en très grande partie le mouvement de la plupart des séries publiées Y_t (autrement, ces séries ne seraient pas publiées), l'élaboration de modèles pour les signaux θ_t peut s'inspirer des cas antérieurs de modélisation des séries Y_t où l'on fait abstraction des erreurs d'échantillonnage. Cela donne à croire que les modèles de signaux feront une place importante aux transformations non linéaires, à l'application de différences et aux fonctions de moyenne de régression. Le logarithme est la transformation non linéaire qui est la plus couramment appliquée aux séries chronologiques et à ce propos, $\log(Y_t)$ nous permet de modéliser $\log(\theta_t)$ grâce à l'équation (2.10), avec les conséquences que l'on sait pour l'erreur d'échantillonnage. Les remarques qui suivent s'appliquent plus particulièrement à l'utilisation de (2.1) mais elles valent aussi pour l'utilisation de (2.10). Bien que l'on pourrait envisager d'autres types de transformations que la transformation logarithmique, celles-là ne permettraient pas, en règle générale, d'exprimer convenablement les séries transformées Y_t en fonction de séries transformées θ_t et de l'erreur d'échantillonnage. Par conséquent, pour la modélisation de nombreuses séries chronologiques, il suffit, semble-t-il, de choisir entre la transformation logarithmique ou pas de transformation du tout.

Dans l'hypothèse où e_t a une moyenne nulle (suivant la propriété d'être non biaisé selon le plan) et qu'il n'est pas nécessaire de lui appliquer des différences, on devra appliquer les mêmes différences à θ_t et à Y_t et celles-ci auront la même fonction de moyenne. On peut souvent modéliser la fonction de moyenne à l'aide d'une fonction de régression linéaire, $\mu_t = \hat{X}'_t \hat{\beta}$, pour un vecteur de variables de régression \hat{X}_t et de paramètres $\hat{\beta}$. On se sert souvent de modèles ARMMI (autorégressifs à moyennes mobiles intégrés) pour représenter les différences devant être appliquées et décrire la structure d'autocovariances des valeurs θ_t , une fois les différences appliquées. Une façon convenable d'élaborer un modèle pour θ_t consiste tout d'abord à modéliser Y_t en faisant abstraction de l'erreur d'échantillonnage, puis à définir un modèle ayant les mêmes termes de régression et le même ordre ARMMI pour θ_t . On peut ensuite estimer les paramètres du modèle de θ_t à l'aide des données de la série Y_t et du modèle élaboré antérieurement pour e_t , les paramètres de ce dernier étant fixes. Par un test de diagnostic, on peut être amené à apporter des modifications au modèle de θ_t . Le modèle ajusté

Nous voyons que l'utilisation de $\hat{\theta}$ dans une perspective de plan engendre une variance moins élevée mais un biais plus grand puisque $\Sigma_e - \text{Var}(\hat{\theta}) - \hat{\theta} \mid \hat{\theta})$ est une matrice semi-positive. Quant à savoir si cette application rend l'EQM conditionnelle (3.3) inférieure à Σ_e , l'EQM de \bar{X} , cela dépend des deux derniers termes de (3.3) et, par conséquent, de $\hat{\theta}$. Il peut y avoir des valeurs de $\hat{\theta}$ pour lesquelles l'EQM conditionnelle de $\hat{\theta}$ est supérieure à Σ_e mais, en règle générale, l'extraction de signal a pour effet de réduire l'EQM d'une valeur équivalant à $\Sigma_e \Sigma_Y^{-1} \Sigma_e$ puisque l'espérance inconditionnelle du terme entre crochets dans l'équation (3.3) est nulle. (Evidemment, l'équation (3.3) est inutilisable en pratique puisqu'elle dépend de $\hat{\theta}$.) De plus, comme nous l'avons mentionné plus tôt, l'erreur de modélisation accroîtra l'EQM de $\hat{\theta}$; cela nous amène à nous poser une autre question fondamentale, à laquelle il est encore plus difficile de répondre (voir Eitinge et Fuller 1989): l'EQM inconditionnelle réelle de $\hat{\theta}$ est-elle inférieure, égale ou supérieure à Σ_e ?

4. CONSIDÉRATIONS RELATIVES À L'APPLICATION

Nous avons vu dans la section 2 que pour appliquer l'approche chronologique à l'estimation dans les enquêtes, il fallait estimer la structure d'autocovariances des erreurs d'échantillonnage de même que la moyenne et la structure d'autocovariances du signal et calculer les valeurs estimées θ'_i et $\text{Var}(\theta'_i - \theta_i)$. Pour ce qui est des opérations d'estimation, on se sert habituellement de modèles chronologiques. Ces opérations sont d'ailleurs analysées dans Bell et Hillmer (1989). La présente section renferme des remarques générales. Nous supposons que Y_i est un estimateur non biaisé selon le plan de θ_i . Dans la section suivante, nous donnons un exemple d'application en nous servant de deux séries chronologiques tirées de la Retail Trade Survey du U.S. Bureau of the Census.

Les autocovariances d'erreurs d'échantillonnage, $\text{Cov}(e_i, e_{i+k})$, peuvent être estimées comme les variances d'échantillonnage, $\text{Var}(e_i)$; l'opération est simple et peut s'effectuer suivant plusieurs méthodes. (Voir Wolter (1985).) Dans la pratique, il peut s'avérer difficile de raccorder des microdonnées d'enquête sur plusieurs périodes afin d'estimer directement les covariances d'erreurs d'échantillonnage. Néanmoins, nous allons supposer ici que nous disposons d'estimations de ce genre, $\text{Cov}(e_i, e_{i+k})$, pour une série de périodes t et de décalages k . Malheureusement, s'il devait y avoir des erreurs d'échantillonnage importantes (circonstances où les méthodes d'analyse chronologique peuvent s'avérer particulièrement utiles), les estimations d'autocovariance afficheraient probablement elles-mêmes une forte variance. Cela donne à penser que l'on pourrait faire une sorte de calcul de moyennes afin d'améliorer ces estimations.

Premièrement, si nous supposons que e_i est stationnaire en covariance, de sorte que $\text{Cov}(e_i, e_{i+k}) \equiv \gamma_e(k)$ dépend de k mais non de i , alors chaque valeur $\text{Cov}(e_i, e_{i+k})$ est une estimation de $\gamma_e(k)$ et il suffit de faire la moyenne de ces valeurs, c.-à-d. calculer $\hat{\gamma}_e(k) = (T - k)^{-1} \sum_i \text{Cov}(e_i, e_{i+k})$ si nous connaissons $\text{Cov}(e_i, e_{i+k})$ pour $i = 1, \dots, T - k$. Par ailleurs, on peut calculer la moyenne de $\text{Cov}(e_i, e_{i+k}) = \text{Cov}(e_i, e_{i+k}) / [\text{Var}(e_i) \text{Var}(e_{i+k})]^{1/2}$ par rapport à t pour estimer $\text{Corr}(e_i, e_{i+k})$, qui dépend aussi de k mais non de i , et on peut, comme avant, faire la moyenne des estimations de la variance. Supposons maintenant que e_i est stationnaire en covariance relative, de sorte que $\text{Cov}(e_i/\theta_i, e_{i+k}/\theta_{i+k}) = \text{Cov}(u_i, u_{i+k}) \equiv \gamma_u(k)$ dépend de k mais non de i . Si u_i est $O_p(r_i)$ pour tous i , comme dans la section 3.2, alors d'après l'équation (3.2) et le théorème 6.2.5 de Wolter (1985), $\text{Cov}(\log(u_i), \log(u_{i+k})) = \text{Cov}(u_i, u_{i+k}) + O_p(r_i^3) \approx \gamma_u(k)$. Considérant que $\text{Cov}(e_i, e_{i+k})$ est une estimation de $\text{Cov}(u_i, u_{i+k})$, on peut faire la moyenne de ces estimations par rapport à t pour estimer $\gamma_u(k)$. Par ailleurs, en nous servant du corollaire 5.1.5 de Fuller (1976), nous pouvons montrer que $\text{Corr}(\log(u_i), \log(u_{i+k})) = \text{Corr}(u_i, u_{i+k}) + O_p(r_i^3)$, et en considérant que $\{\text{Cov}(e_i, e_{i+k})/Y_i Y_{i+k}\} / \{[\text{Var}(e_i) \text{Var}(e_{i+k})]^{1/2} / Y_i Y_{i+k}\} = \text{Corr}(e_i, e_{i+k})$ donne une estimation de $\rho^n(k) \equiv \text{Corr}(u_i, u_{i+k})$, nous pouvons faire la

Résultat 3.5: Si Y_t est non biaisé selon le plan pour tous t , alors θ_t et e_t sont des séries chronologiques non corrélées.

Démonstration: Considérons $\text{Cov}(\theta_t, e_j)$ pour deux périodes quelconques t et j . Comme X_j est non biaisé selon le plan, $E(e_j | \tilde{Q}) = E(X_j - \theta_j | \tilde{Q}) = 0$ et par conséquent, $E[E(e_j | \tilde{Q})] = E(e_j) = 0$. De plus, $E(\theta_t \cdot e_j | \tilde{Q}) = \theta_t \cdot E(e_j | \tilde{Q}) = 0$, ce qui implique que $E(\theta_t \cdot e_j) = 0$. Donc, $\text{Cov}(\theta_t, e_j) = E(\theta_t \cdot e_j) - E(\theta_t)E(e_j) = 0$.

Remarque: Si $E(e_j | \tilde{Q})$ ne dépend pas de \tilde{Q} , alors on dira que e_j est indépendante de \tilde{Q} , "en espérance", ce qui est plus fort que l'absence de corrélation entre e_j et \tilde{Q} mais moins fort que l'indépendance stochastique (sauf s'il y a normalité). En réalité, il suffit que $E(e_t | \tilde{Q}) = E(X_t | \tilde{Q}) - \theta_t$ ne dépende pas de \tilde{Q} pour que θ_t et e_t soient des séries non corrélées. On verra cela dans les cas où X_t comme estimateur de θ_t , est entaché d'un biais additif constant (ne dépendant pas de \tilde{Q}_t) ou (si l'on se fonde sur le résultat 3.6 ci-dessous) d'un biais en pourcentage (multiplicatif) constant.

Considérons maintenant la formule de décomposition logarithmique (2.10) pour le cas où les Y_t sont non biaisés selon le plan. Nous supposons que \tilde{u}_j est représenté par $O_p(r_t)$, où $r_t \rightarrow 0$ lorsque $t \rightarrow \infty$, comme dans le modèle de superpopulation de la section précédente; toutefois, pour des raisons de commodité, nous omettons ici l'indice supérieur t des variables aléatoires. (Voir Wolter (1985, p. 222) pour une définition du degré dans le symbole de probabilité $O_p(r_t)$. Par exemple, pour l'estimation d'une moyenne de population, nous aurions souvent $\text{Var}(\tilde{u}_j) \leq K/n_j$, où K est une constante et n_j la taille de l'échantillon au temps j pour la population t . Alors, d'après Wolter (1985, théorème 6.2.1), $\tilde{u}_j = O_p(n_j^{-1/2})$. En appliquant la méthode de linéarisation de Taylor à $\log(u_j) = \log(1 + \tilde{u}_j)$, nous obtenons (d'après Wolter (1985, théorème 6.2.5

$$\log(u_j) = \tilde{u}_j + O_p(r_t^2). \quad (3.2)$$

Cette équation nous amène au résultat suivant:

Résultat 3.6: Si Y_t est non biaisé selon le plan pour tous t et \tilde{u}_j est $O_p(r_t)$, alors les termes $O_p(r_t^3)$, $\log(\theta_t)$ et $\log(u_t)$ représentent des séries chronologiques non corrélées.

Démonstration: D'après le théorème 6.2.5 de Wolter (1985), $\text{Cov}(\log(\theta_t), \log(u_j)) = \text{Cov}(\log(\theta_t), \tilde{u}_j) + O_p(r_t^3)$. Notons que $E(\tilde{u}_j | \tilde{Q}) = E(e_j | \tilde{Q})/\theta_j = 0$ suppose que $E(\tilde{u}_j) = 0$, et $E(\log(\theta_t)\tilde{u}_j | \tilde{Q}) = \log(\theta_t)E(\tilde{u}_j | \tilde{Q}) = 0$ suppose que $E(\log(\theta_t)\tilde{u}_j) = 0$; par conséquent, $\text{Cov}(\log(\theta_t), \tilde{u}_j) = 0$, ce qu'il fallait démontrer.

3.3 Propriétés de plan des estimations d'extraction de signal

Inconditionnellement, l'estimateur $\tilde{\theta}$ définie en (2.3) est non biaisé ($E(\tilde{\theta}) = E(\tilde{\theta}) = \tilde{\theta}$) et a une EQM minimum telle que définie en (2.6). Lorsqu'on envisage cet estimateur par rapport à un plan, il est facile de constater que ces propriétés ne sont plus aussi vraies. Supposons que nous avons des estimateurs \tilde{X} non biaisés selon le plan, c.-à-d. $E(\tilde{X} | \tilde{Q}) = \tilde{\theta}$. D'après (2.2) et (2.4) nous avons $\tilde{\theta} - \tilde{\theta} = (I - \Sigma_e \Sigma_Y^{-1}) \tilde{e} - \Sigma_e \Sigma_Y^{-1} (\tilde{\theta} - \tilde{\theta})$. Par quelques opérations algébriques, nous pouvons exprimer le biais, la variance et l'EQM de plan de $\tilde{\theta}$:

$$\begin{aligned} E(\tilde{\theta} | \tilde{Q}) - \tilde{\theta} &= -\Sigma_e \Sigma_Y^{-1} (\tilde{\theta} - \tilde{\theta}), \\ \text{Var}(\tilde{\theta} | \tilde{Q}) &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e \Sigma_Y^{-1} \Sigma_e, \\ E[(\tilde{\theta} - \tilde{\theta})(\tilde{\theta} - \tilde{\theta})'] &= \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e \end{aligned} \quad (3.3)$$

$$1 \leq \exp[E[\log(u_i')^2]] \leq E[\exp[\log(u_i')^2]] = E[(u_i')^2].$$

Or, $E[(u_i')^2] = \text{Var}(u_i') + [E(u_i')]^2 \rightarrow 1$, de sorte que $\exp[E[\log(u_i')^2]] \rightarrow 1$ et, partant, $E[\log(u_i')^2] \rightarrow 0$. Nous obtenons ainsi $\text{Var}(\log(u_i')) \rightarrow 0$, le résultat recherché.

Nous pourrions aussi observer une convergence en probabilité en fixant une borne pour $\log(u_i')$. Étant donné que $\log(\theta_i')$ est un estimateur de $\log(\theta_i)$, nous formulons ci-dessous un corollaire du Résultat 3.3 concernant l'utilisation de $\exp[\log(\theta_i')]$ comme estimateur de θ_i .

Corollaire 3.4: Si Y_i' converge en moyenne quadratique vers θ_i' lorsque $\ell \rightarrow \infty$, pour $i = 1, \dots, T$, alors (voir (2.11)) $\exp[\log(\theta_i')]$ converge en probabilité vers θ_i' lorsque $\ell \rightarrow \infty$, pour $i = 1, \dots, T$.

Démonstration: Puisque la convergence en moyenne quadratique de $\log(\theta_i')$ vers $\log(\theta_i)$ implique qu'il y a aussi convergence en probabilité, la démonstration du corollaire est faite étant donné que $\exp(\cdot)$ est une fonction continue (Chung 1968, p. 66).

Nous arrivons évidemment aux mêmes conclusions pour ce qui a trait à l'utilisation de $\exp[\log(\theta_i')] + \text{Var}(\log(\theta_i')) - \log(\theta_i')$ comme estimateur de θ_i , car $\text{Var}(\log(\theta_i')) \rightarrow 0$ lorsque $\ell \rightarrow \infty$.

3.2 Absence de corrélation entre θ et e

Dans sa forme classique, l'extraction de signal de séries chronologiques, représentée par les équations (2.3) à (2.8), suppose invariabilité que θ_i' et e_i' sont non corrélées à toutes les avances et à tous les retards (suivant l'hypothèse de la normalité, nous dirions que ces séries sont indépendantes l'une de l'autre). Dans les articles antérieurs sur l'approche chronologique appliquée à l'estimation dans les enquêtes à passages répétées, on se limite à supposer la même chose. Or, il n'est pas sûr que cette hypothèse soit fondée car θ_i' et e_i' dépendent des mêmes unités de population. Heureusement, nous pouvons établir qu'elle est juste suivant des conditions assez générales. (Tam (1987) explique comment une telle hypothèse peut ne plus être valide dans une approche fondée explicitement sur un modèle.)

Posons y_{it}'' comme la valeur de la caractéristique étudiée pour l'unité i de la population au temps t , et $\Omega_i' = \{y_{it}'' : i = 1, \dots, N_i\}$ comme la série de valeurs pour les N_i unités. Considérons les périodes $t = 1, \dots, T$ et posons $\tilde{\Omega} = (\Omega_1, \dots, \Omega_T)$. Les y_{it}'' sont des variables aléatoires, tout comme $\theta_i' = \theta_i(\Omega_i')$, qui est une fonction des y_{it}'' . L'échantillon au temps t , s_t' (ceci désignant l'indice et non la valeur des unités choisies), a une probabilité de sélection $p(s_t' | \tilde{\Omega})$. L'estimateur X_t' de θ_i' est une fonction des valeurs y_{it}'' pour les unités échantillonnées, donc fonction de Ω_i' et de s_t' , c.-à-d. $X_t' = X_t'(\Omega_i', s_t')$. Nous pourrions faire en sorte que X_t' dépende d'échantillons rapportant à d'autres périodes que t , mais nous nous en abstiendrons pour des raisons de simplicité.

Nous considérons des estimateurs X_t' de θ_i' qui sont *non biaisés selon le plan* et que nous définissons $E(X_t' | \tilde{\Omega}) = \sum s_t' X_t' p(s_t' | \tilde{\Omega}) = \theta_i'$. Nous pourrions exprimer aussi bien ces estimateurs par la formule $E(X_t' | \Omega_i') = \sum s_t' X_t' p(s_t' | \Omega_i') = \theta_i'$, il faudrait alors supposer que le processus d'échantillonnage est tel que $p(s_t' | \tilde{\Omega}) = p(s_t' | \Omega_i')$, de sorte que $E(X_t' | \tilde{\Omega}) = E(X_t' | \Omega_i')$. Si le plan d'échantillonnage est non informatif, s_t' et $\tilde{\Omega}$ sont indépendants, ce qui implique que $p(s_t' | \tilde{\Omega}) = p(s_t' | \Omega_i') = p(s_t')$ et les deux formules ci-dessus se ramènent à $\sum s_t' X_t' p(s_t') = \theta_i'$. Cette dernière formule correspond à la définition habituelle, qui suppose en règle générale que les y_{it}'' , et par voie de conséquence Ω_i' et θ_i' , sont fixes. (L'hypothèse $p(s_t' | \tilde{\Omega}) = p(s_t' | \Omega_i')$ fait que le processus d'échantillonnage au temps t ($p(s_t' | \tilde{\Omega})$) dépend des valeurs de la population au temps t (Ω_i') et suppose que les valeurs de la population pour d'autres périodes que t (Ω_j' pour $j \neq t$) n'apportent aucune nouvelle information à propos de s_t' . Ce genre de situation peut se produire lorsque l'échantillonnage se fait avec une probabilité proportionnelle à la taille d'une variable auxiliaire au temps t et que cette variable auxiliaire est en corrélation avec les valeurs y_{it}'' uniquement à cette période.) Les hypothèses avancées ici pourraient, à la rigueur, être généralisées.

Le premier terme du membre de droite converge en moyenne quadratique vers $\hat{\theta}$; le second terme du membre de droite a pour moyenne $\hat{\theta}$ et pour variance $\sum_e^t (\sum_e^t)^{-1} \sum_e^t$, qui converge vers 0 lorsque $t \rightarrow \infty$. Comme les deux termes convergent en moyenne quadratique vers $\hat{\theta}$, il en est de même pour $\hat{\theta}_t^* - \hat{\theta}^*$.

La convergence en probabilité est une notion plus connue en théorie des sondages. Si Y_t^i converge en probabilité vers θ_t^i lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$ et qu'il existe des variables aléatoires ξ_t^i ayant une variance finie de telle sorte que $|Y_t^i - \theta_t^i| \leq \xi_t^i$ uniformément en t , alors θ_t^i converge en probabilité de Y_t^i vers θ_t^i implique une convergence en moyenne quadratique de Y_t^i vers θ_t^i (Chung 1968, p. 64). Par conséquent, nous établissons ce qui suit à l'aide du Résultat 3.1:

Résultat 3.2: Si Y_t^i converge en probabilité vers θ_t^i lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$ et qu'il existe des variables aléatoires ξ_t^i ayant une variance finie de telle sorte que $|Y_t^i - \theta_t^i| \leq \xi_t^i$ uniformément en t , alors θ_t^i converge en probabilité vers θ_t^i lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$.

Les résultats ci-dessus montrent que si l'erreur rattachée aux estimations originales Y_t^i de θ_t^i est faible (\sum_e^t est petite), l'écart $\theta_t^i - \theta_t^*$ sera aussi faible. Cela vient de ce que, dans l'équation (3.1), $\hat{\theta} - \bar{X}$ s'amenuise lorsque \sum_e^t s'amenuise; ainsi, lorsque les estimations originales Y_t^i sont entachées d'une faible erreur, l'approche chronologique n'a pas vraiment d'effet sur ces estimations. Binder et Dick (1986) ont noté le phénomène et précisé que, dans un tel cas, le modèle chronologique utilise importe peu. Autrement dit, (3.1) convergera vers $\hat{\theta}$ si $\sum_e^t \rightarrow 0$, et la condition de convergence n'a rien à voir avec $\hat{\theta}$ ou \sum_e^t . En vertu des résultats ci-dessus, nous pouvons donc remplacer $\hat{\theta}$, \sum_e^t et \sum_e^t par les estimations $\hat{\theta}_t^i$, \sum_e^t et \sum_e^t (qui viennent le plus souvent de modèles estimés - voir sections 4 et 5), à la condition que $\hat{\theta}_t^i$ et \sum_e^t convergent vers une valeur quelconque lorsque $t \rightarrow \infty$ (la valeur a peu d'importance, pourvu que la limite de \sum_e^t soit définie positive) et que $\sum_e^t \rightarrow 0$, ce qui devrait être le cas si $\sum_e^t \rightarrow 0$. En outre, il est clair que les résultats ci-dessus valent pour les séries non stationnaires, où $\hat{\theta}$ est défini par (2.7) au lieu de (2.4). Autant l'approche chronologique donne peu de résultats lorsque les estimations originales Y_t^i sont entachées d'une faible erreur, autant elle est profitable dans le cas contraire - c'est-à-dire lorsque $\text{Var}(e_t^i)$ est élevée.

Nous pouvons aussi étudier la convergence d'estimateurs dans les cas où nous utilisons des logarithmes et estimons $\log(\theta_t^i)$ au moyen de l'équation (2.11). Posons alors $\sum_e^t = \text{Var}(\log(\hat{\theta}_t^i))$ où $\log(\hat{\theta}_t^i) = (\log(u_t^i), \dots, \log(u_t^T))'$. Si nous utilisons les logarithmes, il est raisonnable de supposer que Y_t^i et θ_t^i ne peuvent avoir de valeur nulle, c'est-à-dire $|Y_t^i| \geq \kappa$ et $|\theta_t^i| \geq \kappa$ (avec quasi-certitude) pour tous t et tous i pour une constante $\kappa > 0$.

Résultat 3.3: Si Y_t^i converge en moyenne quadratique vers θ_t^i lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$, alors $\log(Y_t^i)$ converge en moyenne quadratique vers $\log(\theta_t^i)$ et $\log(\hat{\theta}_t^i)$ converge également en moyenne quadratique vers $\log(\theta_t^i)$ lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$.

Démonstration: L'équivalence logarithmique de (3.1) est

$$\log(\hat{\theta}_t^i) - \log(\theta_t^i) = (\log(\bar{X}_t^i) - \log(\hat{\theta}_t^i) - \log(\bar{X}_t^i) - 1) \log(\bar{X}_t^i) - (\bar{X}_t^i - \bar{\theta}_t^i).$$

Si nous pouvons montrer que $\sum_e^t \rightarrow 0$, la démonstration sera faite car cette convergence implique que $\log(\bar{X}_t^i)$ converge en moyenne quadratique vers $\log(\theta_t^i)$, et le second terme du membre de droite de l'équation suit la même tendance que le terme correspondant de l'équation (3.1). Notons que

$$E[(u_t^i)^2] = E[(e_t^i)^2 / (\theta_t^i)^2] \leq E[(e_t^i)^2] / \kappa^2 \rightarrow 0 \text{ lorsque } t \rightarrow \infty,$$

par conséquent, $E[(u_t^i)^2] = E[(u_t^i - 1)^2] \rightarrow 0$. Donc, $\text{Var}(u_t^i) \rightarrow 0$ et $E(u_t^i) \rightarrow 1$. D'après l'inégalité de Jensen (Chung 1968, p. 45), puisque $\exp(\cdot)$ est une fonction convexe,

Hausman et Watson 1985; Pfeiffermann 1991; l'estimation de la tendance d'une série chronologique (et le problème connexe de la constatation d'une variation statistiquement significative à long terme) (Smith 1978); l'étalement, c'est-à-dire le rapprochement des résultats d'une enquête à passages répétées et de ceux d'une autre enquête ou d'un recensement qui sert à estimer les mêmes caractéristiques de la population (Hillmer et Trabelsi 1987; Trabelsi et Hillmer 1990); et l'inférence concernant les propriétés chronologiques de la série θ_t , réelle par rapport aux modèles économiques (Bell et Wilcox 1990).

Enfin, notons que les erreurs non dues à l'échantillonnage ne sont pas prévues dans la formule de décomposition (2.1) ou (2.10) et qu'elles ne sont pas considérées de façon explicite dans l'approche chronologique. On ne peut encore dire si ce genre d'erreurs posent habituellement un problème plus sérieux ou moins sérieux pour l'approche chronologique que pour l'approche classique, mais il serait intéressant de connaître les effets que peuvent avoir les erreurs non dues à l'échantillonnage (réelles ou prévues) sur les estimateurs de séries chronologiques lorsque ceux-ci sont utilisés dans des conditions précises.

3. CONSIDÉRATIONS THÉORIQUES

Dans cette section, nous présentons des résultats théoriques concernant l'approche chronologique et exposons certaines propriétés des estimateurs pertinents.

3.1 Convergence des estimateurs de séries chronologiques

Suivant Fuller et Isaki (1981), nous définissons Y_t^i (d'après l'échantillon i au temps t) comme une série d'estimateurs de la caractéristique θ_t^i de la population i au temps t , où les populations et les échantillons pour $t = 1, 2, \dots$ sont emboîtés. (Voir l'article précité pour plus de détails.) Définissons $\tilde{Y}_t^i, \tilde{\theta}_t^i, \tilde{e}_t^i, \tilde{\mu}_t^i, \Sigma_{\theta}^i, \Sigma_{\tilde{\theta}}^i, \Sigma_{\tilde{e}}^i, \Sigma_{\tilde{\mu}}^i$, et θ_t^i de la manière habituelle. Nous allons voir ce que deviennent les estimateurs de séries chronologiques $\tilde{\theta}_t^i$ lorsque les estimateurs \tilde{Y}_t^i sont convergents, c.-à-d. $Y_t^i \rightarrow \theta_t^i$ d'une manière quelconque lorsque $t \rightarrow \infty$ pour $i = 1, \dots, T$, la longueur de la série (T) étant constante. Pour le moment, nous supposons que μ_t^i , Σ_{θ}^i , et $\Sigma_{\tilde{\theta}}^i$ sont connues pour chaque i , ce qui signifie généralement que les modèles chronologiques (y compris les valeurs des paramètres) pour les composantes sont connus. Puisque μ_t^i et $\Sigma_{\tilde{\theta}}^i$ sont en réalité des paramètres de superpopulation pour la série chronologique que nous voulons estimer (θ_t^i), nous supposons que chacun d'eux est unique pour toutes les populations i , c'est-à-dire $\mu_t^i \equiv \tilde{\mu}_t^i$ et $\Sigma_{\theta}^i \equiv \Sigma_{\tilde{\theta}}^i$ (une matrice définie positive) pour tous i . Nous posons aussi cette hypothèse en partie pour des raisons de commodité, puisque nous obtenons les mêmes résultats en supposant que $\tilde{\mu}_t^i \rightarrow \mu_t^i$ et $\Sigma_{\tilde{\theta}}^i \rightarrow \Sigma_{\theta}^i$ lorsque $t \rightarrow \infty$.

D'après l'équation (2.5), il semble que $\tilde{Y}_t^i \rightarrow \theta_t^i$ implique que $\tilde{\theta}_t^i \rightarrow \theta_t^i$ pourvu que $\Sigma_{\tilde{e}}^i \rightarrow 0$. Cette condition nous porte à croire que Y_t^i doit converger en moyenne quadratique vers θ_t^i . Nous considérons donc les estimateurs Y_t^i de θ_t^i tels que $E[(Y_t^i - \theta_t^i)^2] = E[(e_t^i)^2] \rightarrow 0$ lorsque $t \rightarrow \infty$. Puisque $E[(e_t^i)^2] = \text{Var}(e_t^i) + [E(e_t^i)]^2$ cela implique que $\text{Var}(e_t^i) \rightarrow 0$ et $E(e_t^i) \rightarrow 0$ également. Le fait de supposer que Y_t^i converge en moyenne quadratique vers θ_t^i , pour $t = 1, \dots, T$, implique donc que $\Sigma_{\tilde{e}}^i \rightarrow 0$. Nous pouvons maintenant établir ce qui suit:

Résultat 3.1: Considérons l'estimateur $\tilde{\theta} = (\theta_1, \dots, \theta_T)'$ défini en (2.4). Si Y_t^i converge en moyenne quadratique vers θ_t^i lorsque $t \rightarrow \infty$, pour $i = 1, \dots, T$, alors θ_t^i converge en moyenne quadratique vers θ_t^i lorsque $t \rightarrow \infty$, pour $t = 1, \dots, T$.

Démonstration: D'après l'équation $\tilde{Y}_t^i = \theta_t^i + \tilde{e}_t^i$ où $\Sigma_{\tilde{e}}^i \rightarrow 0$, nous avons $\Sigma_{\tilde{Y}}^i \rightarrow \Sigma_{\theta}^i$ (même si $\tilde{\theta}^i$ et \tilde{e}^i sont corrélés). D'après (2.4), nous avons

$$\tilde{\theta}^i - \theta^i = (\tilde{Y}^i - \theta^i)' (\Sigma_{\tilde{Y}}^i)^{-1} (\tilde{Y}^i - \theta^i). \tag{3.1}$$

Notons que pour résoudre les équations (2.3) à (2.6), il faut connaître $\hat{\mu}$ ainsi que deux des trois matrices de covariances Σ_Y , Σ_θ et Σ_ε (la troisième pouvant être déterminée à l'aide de l'équation (2.2)). En pratique, on ne connaîtra pas précisément ces matrices et il faudra donc les estimer. Par conséquent, il n'est pas possible de connaître avec exactitude l'estimateur linéaire à erreur quadratique moyenne minimum $\hat{\theta}$ et les équations (2.6) ou (2.8) produisent une valeur sous-estimée de l'erreur quadratique moyenne (EQM) puisqu'elles ne tiennent pas compte des erreurs de modélisation. (Voir Binder et Dick (1989) et Elling et Fuller (1989).) L'hypothèse fondamentale qui sert de base à l'application des équations ci-dessus est que l'on peut estimer convenablement $\hat{\mu}$ et Σ_Y à l'aide des données de la série X_t et d'un modèle chronologique, et tout aussi convenablement Σ_ε à l'aide de micro-données d'enquête et des données du plan de sondage (et probablement aussi à l'aide d'un modèle). Nous examinons plus en détail cette question dans la section 4 et illustrons l'approche chronologique dans la section 5.

2.2 Considérations générales sur l'approche chronologique

Smith (1978), Jones (1980) et Binder et Dick (1986) analysent l'approche dite de l'estimation linéaire non biaisée à variance minimum (ELNVM). Bien que l'ELNVM et l'approche chronologique permettent toutes deux d'estimer θ_t à l'aide de données se rapportant à d'autres périodes que t , elles se distinguent l'une de l'autre en ceci que, dans la première, on tient les valeurs θ_t pour fixes et ne considère qu'une seule forme de variation, soit celle due à l'échantillonnage. L'ELNVM est utilisée spécialement dans les cas où il existe plus d'un estimateur direct de θ_t pour chaque période t et où les ε_t sont corrélées d'une période à l'autre à cause de la participation répétée de certaines unités d'échantillonnage (c'est le cas notamment de nombreuses enquêtes par panel avec renouvellement). Ainsi, les résultats des moindres carrés généralisés et la corrélation des ε_t font que X_j , pour $j \neq t$, sert à estimer θ_t . Nous pouvons voir une différence par rapport à l'équation (2.1), où il existe un seul estimateur direct (X_t) de θ_t , en ce sens que dans ce dernier cas, on obtient l'ELNVM en posant $\text{Var}(\theta_t) \rightarrow \infty$. Alors, de $\theta_t^{-1} \rightarrow 0$ et l'équation (2.5) devient $\hat{\theta} = \tilde{X}$, de sorte que s'il n'existe pas plusieurs estimateurs de θ_t , l'ELNVM n'utilise que X_t pour estimer θ_t . Ces remarques concernent ordinairement l'estimation composite (Rao et Graham 1964; Wolter 1979), qui est souvent utilisée comme équivalent de l'ELNVM.

Pour ce qui a trait à l'approche chronologique, il est permis de se demander pourquoi considérer θ_t comme une série stochastique? Cette question a été étudiée par SJS et analysée assez longuement par Smith (1978). Deux observations ressortent de ces analyses: 1) dans la modélisation, les utilisateurs de données d'enquêtes à passages répétées considèrent X_t comme une série stochastique et ils envisageraient θ_t de la même façon s'ils en connaissaient les éléments (comme c'est le cas pour les enquêtes où le niveau d'erreur est très faible); 2) dans les textes classiques (par ex.: Patterson, 1950) sur l'estimation dans les enquêtes à passages répétées (ELNVM), on suppose l'existence d'une structure chronologique pour chacune des unités de la population tout en continuant d'affirmer que θ_t , qui est une fonction de ces unités (par ex.: le total), est une série de valeurs fixes n'ayant aucun rapport entre elles. À vrai dire, si nous supposons que θ_t est une suite de valeurs fixes qui n'ont aucun lien entre elles, alors les données relatives à chacune des périodes n'ont aucun rapport avec le comportement futur de la série réelle θ_t . Si tel était le cas, on pourrait s'interroger sur l'utilité même d'une enquête car les données ne seraient pas aussitôt publiées qu'elles seraient périmees. En réalité, il s'agit de savoir si, oui ou non, nous pouvons estimer la structure chronologique de θ_t et de ε_t suffisamment bien pour que cela nous soit profitable dans le processus d'estimation, si les avantages en question sont notables et si l'opération comporte des inconvénients et lesquels?

Outre qu'elle fournit l'occasion d'améliorer l'estimation dans les enquêtes à passages répétées, l'approche chronologique peut être utile dans les cas où on ne reconnaît généralement qu'une des deux sources de variation possibles. Elle pourrait bien aussi servir d'approche globale pour tous ces cas. Mentionnons notamment l'estimation provisoire dans les enquêtes à passages répétées (Rao, Srinath et Quenneville 1989); la désaisonnalisation (Wolter et Monsour 1981;

Si on pose par hypothèse la normalité de $(\tilde{\theta}, \tilde{X})$ les équations (2.3) à (2.5) permettent d'obtenir $F(\tilde{\theta} | \tilde{X})$, l'espérance conditionnelle de $\tilde{\theta}$ étant donné \tilde{X} , et l'équation (2.6), la variance conditionnelle, $\text{Var}(\tilde{\theta} | \tilde{X})$.

S'il faut appliquer des différences à Y_t , les résultats précédents ne tiennent plus. Supposons qu'il n'est pas nécessaire d'appliquer des différences à e_t mais qu'il faut appliquer une différence d'ordre un à θ_t et à Y_t (au moyen du facteur $1 - B$ ou $BY_t = X_{t-1}$). Soit la nouvelle série de données $W_t = (1 - B)Y_t = (1 - B)\theta_t + (1 - B)e_t$ pour $t = 2, \dots, T$. Définissons $\Delta = [\Delta_{ij}]$ comme la matrice de différences $(T - 1) \times T$ où $\Delta_{it}'' = -1$, $\Delta_{it+1}'' = 1$, et tous les autres éléments sont nuls, et écrivons $\Delta\tilde{X} \equiv \tilde{W} = \Delta\tilde{S} + \Delta\tilde{e}$. Nous utilisons ensuite

(2.7)
$$\tilde{\theta} = \tilde{X} - \tilde{e} = \tilde{X} - \Sigma^e \Delta' \Sigma^{W-1} \Delta(\tilde{X} - \tilde{w}),$$

(2.8)
$$\text{Var}(\tilde{\theta}) = \Sigma^e - \Sigma^e \Delta' \Sigma^{W-1} \Delta \Sigma^e.$$

Les équations (2.7) et (2.8) sont aussi utiles lorsqu'il faut appliquer un opérateur plus général (par ex. : différences saisonnières) à θ_t et à Y_t (la matrice des différences Δ étant définie en conséquence), pourvu qu'il ne soit pas nécessaire d'appliquer des différences à e_t . Ces équations sont analogues à (2.4) et à (2.6) sauf que $\Delta' \Sigma^{W-1} \Delta$ est substitué à Σ^{Y-1} . Nous trouvons ces équations dans Bell et Hillmer (1990), qui les analysent sur le plan de l'optimalité. Jones (1980) avait exposé essentiellement les mêmes résultats mais sans les expliquer vraiment.

Scott et Smith (1974) et SSJ ont utilisé la méthode classique d'extraction de signal (c.-à-d. les équations (2.3) à (2.6)) mais en se fondant sur des fonctions génératrices de covariances plutôt que des matrices de covariances. Bell (1984) utilise la même méthode pour des modèles prévoyant l'application de différences. Une autre approche (Binder et Dick 1989; Bell et Hillmer 1989) consiste à exprimer des modèles chronologiques de θ_t et de e_t sous forme de modèles d'espace d'états puis à utiliser le filtre de Kalman, façon efficace, semble-t-il, d'obtenir les matrices définies ci-dessus. En outre, Tam (1987) se sert du filtre de Kalman dans une approche fondée explicitement sur un modèle pour l'analyse des données d'enquêtes à passages répétés. Dans le reste de cet article, il sera le plus souvent question des équations (2.3) à (2.6); cependant, nos commentaires valent tout aussi bien pour les équations (2.7) et (2.8).

Il arrive souvent que, pour des séries chronologiques Y_t et θ_t , toujours positives, l'on veuille calculer le logarithme de Y_t pour faire de θ_t et de e_t des séries stationnaires. Dans un tel cas, on reformule (2.1) comme suit:

(2.9)
$$Y_t = \theta_t(1 + u_t), \quad e_t u_t = e_t / \theta_t \text{ et } u_t = 1 + \tilde{u}_t.$$

En utilisant les logarithmes, on obtient

(2.10)
$$\log(Y_t) = \log(\theta_t) + \log(1 + \tilde{u}_t) = \log(\theta_t) + \log(u_t).$$

En posant que \tilde{w} et $\Sigma^{\tilde{\theta}}$ représentent désormais $\log(\tilde{\theta}) \equiv (\log(\theta_1), \dots, \log(\theta_T))'$, et que $\Sigma^Y = \Sigma^{\tilde{\theta}} + \Sigma^u$ représente $\log(\tilde{Y})$, la formule de l'estimateur, équivalente à (2.4), est

(2.11)
$$\log(\tilde{\theta}) = \tilde{w} + [I - \Sigma^u \Sigma^{Y-1}] (\log(\tilde{Y}) - \tilde{w}).$$

Les formules équivalentes aux équations (2.6) à (2.8) sont évidentes. Pour estimer θ_t , on utilise $\exp[\log(\theta_t)]$; par ailleurs, on pourrait utiliser $\exp[\log(\theta_t) + \text{Var}(\log(\theta_t)) / 2]$ pour obtenir une estimation "moins biaisée" de θ_t , qui aurait une erreur quadratique moyenne minimum (voir Granger et Newbold 1976).

Dans le cas des estimateurs classiques, on parle uniquement de variation due à l'échantillonnage – du fait que l'on n'a pas observé toutes les unités de la population. Dans le cas de l'analyse chronologique, on parle de variation due au fait que l'on ne peut définir parfaitement (le plus souvent de façon linéaire) une série chronologique à l'aide de données antérieures. Considérons la formule de décomposition:

$$(2.1) \quad Y_t = \theta_t + e_t,$$

où Y_t est une estimation d'enquête au temps t , θ_t est la valeur de la caractéristique de la population au temps t et e_t est l'erreur d'échantillonnage. La variabilité d'échantillonnage de e_t est le point de mire des méthodes d'estimation classiques, qui considèrent θ_t comme fixe. En revanche, l'approche chronologique prévoit que Y_t , θ_t , et e_t peuvent toutes les trois varier dans le temps pourvu qu'elles soient aléatoires et qu'on ne puisse en prévoir parfaitement la valeur à l'aide de données antérieures. L'analyse chronologique ordinaire porte directement sur Y_t et ne tient pas compte de l'erreur d'échantillonnage comme telle, si ce n'est comme un élément de l'agrégat Y_t . De fait, les analystes de séries chronologiques font comme si la variation d'échantillonnage n'existait pas et que les valeurs réelles étaient observées. Bref, ce qu'il faut surtout retenir au sujet de l'application des méthodes d'analyse chronologique à l'estimation fondée sur les enquêtes, c'est qu'il existe deux sources de variation stochastique qui sont définies, modélisées et estimées différemment l'une de l'autre.

2.1 Résultats de l'extraction du signal

Supposons qu'il existe des estimations d'enquête Y_t pour une série de périodes identifiées $t = 1, \dots, T$. Soit $\tilde{Y} = (Y_1, \dots, Y_T)'$ et définissons de la même façon $\tilde{\theta}$ et \tilde{e} , de sorte que $\tilde{Y} = \tilde{\theta} + \tilde{e}$. En supposant que les estimations Y_t sont non biaisées et que θ_t et e_t sont non corrélées (voir section 3.2)

$$E(\tilde{Y}) = E(\tilde{\theta}) \equiv \tilde{\mu} \equiv (\mu_1, \dots, \mu_T)'$$

$$(2.2) \quad \Sigma_Y = \Sigma_\theta + \Sigma_e,$$

où E désigne l'espérance de la distribution d'échantillonnage et de la distribution du modèle chronologique, et Σ_Y est la matrice des covariances de \tilde{Y} , etc. Dans les expressions ci-dessus, $\tilde{\mu}$ et Σ_θ ont rapport à la structure chronologique de θ_t , qui n'est pas exposée à la variation d'échantillonnage. Lorsqu'on n'a pas besoin d'appliquer des différences aux séries Y_t , θ_t , ou e_t on sait très bien qu'il est possible d'exprimer l'estimateur linéaire à erreur quadratique moyenne minimum de $\tilde{\theta}$ de la façon suivante (en se servant de l'équation (2.2)) puisque

$$\text{Cov}(\tilde{\theta}, \tilde{Y}) = \Sigma_\theta:$$

$$(2.3) \quad \tilde{\theta} = \tilde{\mu} + \Sigma_\theta \Sigma_Y^{-1} (\tilde{Y} - \tilde{\mu})$$

$$(2.4) \quad = \tilde{\mu} + (I - \Sigma_e \Sigma_Y^{-1}) (\tilde{Y} - \tilde{\mu})$$

$$(2.5) \quad = \tilde{\mu} + (I + \Sigma_e \Sigma_{Y-1}^{-1}) (\tilde{Y}_{-1} - \tilde{\mu}).$$

On sait aussi que la variance de l'erreur de cet estimateur est

$$(2.6) \quad \text{Var}(\tilde{\theta} - \theta) = \Sigma_\theta - \Sigma_\theta \Sigma_Y^{-1} \Sigma_\theta = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e.$$

Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques

WILLIAM R. BELL et STEVEN C. HILLMER¹

RÉSUMÉ

Scott et Smith (1974) et Scott, Smith et Jones (1977) ont proposé d'utiliser les résultats de l'extraction de signal afin d'améliorer les estimations tirées des enquêtes à passages répétés; c'est ce qu'on appelle l'approche chronologique à l'estimation dans les enquêtes à passages répétés. Dans cet article, nous examinons les fondements théoriques de cette approche en précisant qu'elle repose sur la reconnaissance de deux sources de variation – variation chronologique et variation d'échantillonnage – et qu'elle peut aussi servir à résoudre d'autres problèmes qui mettent en évidence ces deux sources de variation. Avec cette approche, nous obtenons des résultats théoriques concernant la convergence selon le plan des estimateurs de séries chronologiques et l'absence de corrélation entre la série du signal et la série des erreurs d'échantillonnage. Nous constatons que, dans la perspective d'un plan de sondage, l'approche chronologique engendre, par rapport aux estimateurs classiques, une variance et une erreur quadratique moyenne moins élevées mais un biais plus grand. Nous voyons brièvement comment appliquer cette approche par la modélisation puis, à titre d'exemple, nous prenons la série des ventes au détail des établissements de restauration (ou restaurants) et des débits de boissons, tirée de la Retail Trade Survey du U.S. Bureau of the Census.

MOTS CLÉS: Enquêtes à passages répétés; séries chronologiques; extraction du signal; U.S. Retail Trade Survey.

1. INTRODUCTION

Scott et Smith (1974) et Scott, Smith et Jones (1977) (ci-après désigné SSJ) ont proposé de recourir à l'analyse de séries chronologiques et plus particulièrement à l'extraction de signal afin d'améliorer les estimations tirées des enquêtes à passages répétés. Lorsqu'on connaît la structure de covariances des estimations d'enquête (Y_t) et des erreurs d'échantillonnage correspondantes (e_t) pour une série de périodes, les résultats de l'extraction de signal peuvent servir à déterminer les fonctions linéaires à erreur quadratique moyenne minimum des observations Y_t , ces fonctions étant les estimateurs des caractéristiques de population à estimer pour une série stochastique θ_t . Concrètement, on se sert des résultats de l'extraction de signal en estimant un modèle chronologique pour la série d'observations Y_t et en estimant la structure de covariances de e_t à long terme à l'aide des données du plan de sondage.

Dans la section 2, nous exposons sommairement l'approche chronologique et ses caractéristiques fondamentales. La section 3 renferme des questions d'ordre théorique tandis que la section 4 présente des considérations relatives à l'application de l'approche. Enfin dans la section 5, nous donnons un exemple d'application en nous servant de deux séries chronologiques tirées de la Retail Trade Survey du U.S. Bureau of the Census.

2. L'APPROCHE CHRONOLOGIQUE: PRINCIPES FONDAMENTAUX ET ANALYSE GÉNÉRALE

Les estimateurs de séries chronologiques se distinguent principalement des estimateurs classiques par le fait qu'ils reconnaissent l'existence de deux sources de variation au lieu d'une.

¹ William R. Bell est principal de recherche à la Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, E.-U. et Steven C. Hillmer est professeur à la School of Business, University of Kansas, Lawrence, Kansas 66045, E.-U.

BIBLIOGRAPHIE

- COLLEDGE, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. Dans *Panel Surveys*, (éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: Wiley & Sons.
- DUNCAN, G.J., et KALTON, G. (1987). Issues of Design and Analysis of Surveys Across Time. *Revue Internationale de Statistique*, 55, 97-117.
- ERNST, L. (1989). Weighiting Issues for Longitudinal Household and Family Estimates. In *Panel Surveys*, (éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: Wiley & Sons.
- HANSON, R.H. (1978). *The Current Population Survey: Design and Methodology*, Technical Paper 40. United States Bureau of the Census. Washington, DC.
- LAVRINI, R. (1987). Manipulation of Spatial Objects by a Peano Tuple Algebra, University of Maryland Technical Report CS-TR-1893, College Park, MD.
- PEANO, G. (1908). La Curva di Peano nel Formulatio Mathematico. Dans *Opere Scelte di G. Peano*, 115-116, Vol. I. Edizioni Cremonesi, Roma, 1957.
- PROGRESSIVE GROCER (1989). 56th Annual Report of the Grocery Industry 1989, Vol. 68, No. 4, Part 2, Stamford CT.
- RAO, J.N.K., et GRAHAM, J.R. (1964). Rotation Designs for Sampling on Repeated Occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SAALFELD, A., FIFIELD, S., BROOME, F., et MEIXLER, D. (1988). Area Sampling Strategies and Payoffs using Modern Geographic Information System Technology. Article non-publié, United States Bureau of the Census, Washington, DC.
- SIRKEN, M. (1970). Household Surveys with Multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- WOLTER, K.M. (1979). Composite Estimation in Finite Population. *Journal of the American Statistical Association*, 74, 604-613.
- WOLTER, K.M. (1986). *Introduction to Variance Estimation*. New York: Springer Verlag.
- WOLTER, K.M. et coll. (1976). Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, Alexandria, VA.

Selon cette règle et les données connues (ex: lecture optique ou non), tous les fonds alloués pour l'enquête peuvent être consacrés à l'échantillonnage des commerces avec lecteurs optiques. Il n'est pas nécessaire de consacrer des ressources au sondage des commerces sans lecteurs optiques.

Pour résoudre le cas ii), posons s comme l'échantillon de commerces sélectionné et soit $s_A = s \cap A$ et $s_B = s \cap B$. Par hypothèse, s_A et s_B ne sont connus qu'après l'étape initiale de travail sur le terrain. Evidemment, chacun de ces ensembles varie avec le temps mais toute référence au temps a été supprimée afin de simplifier la notation.

L'échantillon s_A doit être mis à jour conformément aux règles indiquées dans cet article pour les ouvertures et les fermetures. Dans le cas de s_B il faut établir de nouvelles règles. Voici une règle qui traite les commerces de s_B comme non répondants:

Règle 5. Au temps t imputer au commerce $U_{tjb} \in s_B$ la valeur y_{tjb}/x_{At} , où x_{tjb} est la valeur d'une variable auxiliaire pour le commerce U_{tjb} , y_{At} est le total de la variable d'estimation pour l'échantillon s_A et x_{At} est le total correspondant pour la variable auxiliaire. De plus, l'imputation peut se faire par substitution, par appariement "Hot Deck" ou par un autre moyen. Maintenant, en supposant que la série de données est complète, appliquer l'estimateur standard du paramètre d'intérêt. Au moment où U_{tjb} adopte la lecture optique, il doit être soustrait de s_B et ajouté à s_A , et l'estimation doit être effectuée à l'aide de l'estimateur standard appliqué à la série de données complétée.

Selon la règle 5, la taille effective de l'échantillon est réduite en raison de la variance d'imputation liée à y_{tjb} . La substitution permet de conserver une taille d'échantillon supérieure à celle que déterminent les autres règles mais c'est la méthode la plus coûteuse. Toutes les règles nécessitent un travail restreint mais continu sur le terrain pour vérifier si $U_{tjb} \in s_B$ a adopté ou non la lecture optique.

Une des alternatives à la technique des données manquantes est d'étudier les commerces sans lecteurs optiques à l'aide d'une autre méthode de collecte des données. Selon les données à recueillir, cela peut comprendre une vérification des dossiers de l'entreprise ou une entrevue avec le personnel, par téléphone, par courrier ou en personne. Cette méthode est plus précise que la méthode d'imputation, mais elle est plus coûteuse et exige plus de temps, sans compter le fardeau qu'imposent la gestion et le contrôle de deux méthodes de collecte de données.

En dernier lieu, nous traiterons de l'abandon de la lecture optique par les commerces. Ce genre de cas est assez rare et il n'est traité que pour s'assurer de bien couvrir tout le sujet. Posons $U_{tjb} \in s_A$, c'est-à-dire que i est un commerce avec un lecteur optique dans l'échantillon. Notons que U_{tjb} peut être un commerce qui possède un lecteur optique depuis le début de l'enquête ou un commerce qui a adopté la lecture optique après avoir fait partie de l'échantillon comme commerce sans lecteur optique conformément à la règle 5.

Règle 6. Au moment où U_{tjb} abandonne la lecture optique, il doit être soustrait de s_A , ajouté à s_B et traité par les méthodes de données manquantes, comme dans la règle 5. Les formules standard doivent être appliquées à la série de données complète. Pour simplifier le traitement et le travail sur le terrain, la méthode choisie doit être identique à celle choisie pour traiter les commerces qui adoptent la lecture optique.

Dans le cas inhabituel où un commerce utilise la lecture optique de façon intermittente, il faut le traiter en appliquant les règles 5 ou 6 selon le cas, en mettant à jour chaque fois les échan-

tilillons s_A et s_B .

Le deuxième problème est celui de la période qui s'écoule entre la fermeture du commerce et la mise à jour suivante. Ce problème ne se pose que lorsque la fréquence des mises à jour est inférieure à celle de la collecte des données. Si les deux sont faites en même temps, il n'y a pas de problème. Si la mise à jour est moins fréquente que la collecte, il y a deux solutions:

- Éliminer les commerces fermés de l'échantillon au moment où la fermeture est connue (pour être plus précis du point de vue statistique, cela signifie que les disparitions sont conservées dans l'échantillon avec une valeur de zéro.)
- Garder les commerces fermés dans l'échantillon en leur attribuant une valeur jusqu'au moment de la mise à jour suivante.

La solution a) est la plus simple et la plus claire. À part le problème des ouvertures, elle n'est pas biaisée et permet un bon calcul de l'estimateur de variance. En raison du problème des ouvertures, toutefois, cette solution peut avoir un effet négatif sur la capacité de l'échantillon à mesurer les tendances. Si les fermetures se produisent pendant la première semaine d'un cycle, on peut remarquer un léger déclin dans la série des commerces, non pas en raison d'un changement fondamental des conditions économiques, mais simplement parce que l'échantillon tient compte des fermetures et non des ouvertures. La solution b) est une solution à court terme pour la mesure adéquate des tendances. La notion essentielle est qu'en donnant une valeur aux commerces fermés, nous compensons de façon implicite pour toutes les ouvertures qui ont pu se produire depuis le dernier cycle de mise à jour. Cette solution n'est pas particulièrement élégante et il est difficile de la justifier techniquement. Cependant, l'expérience nous montre que l'univers des commerces est stable à court terme. Les fermetures sont souvent associées à des ouvertures ou compensées par celles-ci et la taille nette de la population reste à peu près constante à court terme. Le United States Bureau of the Census a utilisé cette solution dans son enquête sur le commerce de gros avec des cycles de mise à jour trimestriels et une collecte des données mensuelle. Voir Wolter et coll. (1976).

4.3 Commerces systématiquement exclus et commerces adoptant le système de lecture optique

Dans cette dernière section, nous présentons les règles de mise à jour des échantillons dans le cas des commerces qui sont systématiquement exclus de l'échantillon, soit parce qu'ils ne sont pas pourvus d'un système de lecture optique ou que, même s'ils en possèdent un, ils l'utilisent avec si peu de rigueur que les données produites peuvent être fausses et inutilisables, ou encore qu'ils refusent de participer à l'enquête. Nous nous arrêtons plus particulièrement aux commerces qui n'ont pas le système de lecture optique et aux règles de mise à jour des échantillons dans le cas des commerces qui adoptent le système de lecture optique ou qui l'abandonnent; néanmoins, les propos qui suivent peuvent s'appliquer de façon générale à tous les cas d'exclusion systématique. Nous allons définir A comme l'ensemble des commerces avec lecteurs optiques et B , comme l'ensemble des commerces sans lecteurs optiques; $A \cup B$ représente toute la population.

Nous parlerons en premier lieu des commerces qui adoptent le système de lecture optique. Il y a deux cas à considérer: (i) le système utilisé par tous les commerces est connu avant l'échantillonnage et (ii) le système utilisé n'est connu qu'après l'échantillonnage et uniquement pour les commerces sélectionnés.

Le cas i) est relativement simple. Voici la règle à suivre:

Règle 4. Ne pas inclure les commerces n'utilisant pas la lecture optique (B) dans l'échantillon. Ne choisir l'échantillon que parmi les commerces qui utilisent la lecture optique (A). Si un commerce adopte le lecteur optique, le traiter comme un nouveau commerce et le soumettre à l'échantillonnage des nouveaux commerces. Avant la conversion, les commerces n'ayant pas de lecteur optique (B) doivent être traités par imputation ou par une autre technique pour données manquantes.

4.2 Mise à jour pour la fermeture de commerces

Les règles de mise à jour d'un échantillon dans le temps doivent suivre un important principe général. Elles doivent traiter de la même façon les unités sélectionnées et les unités non sélectionnées. Dans le cas de fermeture de commerces, ce principe signifie que toutes les fermetures, celles qui se produisent à l'intérieur comme à l'extérieur de l'échantillon, doivent être traitées de la même façon dans tous les procédés de mise à jour de l'échantillon. Si ce principe n'est pas respecté, les estimateurs seront biaisés et ce biais risque de s'accroître avec le temps. Dans les paragraphes suivants, nous décrivons des procédés de mise à jour de l'échantillon en cas de fermeture de commerces, qui suivent ce principe essentiel. Nous verrons deux cas: i) les fermetures ne sont pas connues pour toute la population et ii) les fermetures sont connues pour toute la population.

Pour le cas i), nous proposons la règle 2:

Règle 2. Toutes les fermetures de commerces dans l'échantillon sont connues. Les commerces fermés doivent demeurer dans l'échantillon mais leur valeur doit être de 0 (c.-à-d. $y = 0$) au moment de la mise à jour.

Cette règle permet de faire une estimation non biaisée des totaux de population. Les fermetures peuvent faire augmenter la variance de l'estimateur, et les estimateurs de variance vont refléter cette augmentation si les commerces fermés sont conservés dans l'échantillon avec une valeur de 0.

Pour le cas ii), nous proposons la règle 3:

Règle 3. Retirer tous les commerces fermés de la population au moment de la mise à jour suivante. Seuls les commerces ouverts seront soumis à l'échantillonnage, y compris les nouveaux commerces.

La règle 3 fait varier le nombre de commerces B_{ij} dans les segments où il y a eu des fermetures, à moins que le nombre de nouveaux commerces soit égal au nombre de fermetures. Un commerce de remplacement sera choisi à l'intérieur d'un segment dès que le commerce échantillonné de ce segment fermera ses portes – à moins qu'il se produise une fermeture sans qu'un seul commerce ne soit créé, et que $B_{ij} = 0$. Un commerce de remplacement peut être choisi même si le commerce échantillonné est encore ouvert.

Si, exceptionnellement, $B_{ij} = 0$, la taille de l'échantillon diminue de 1. Il serait intéressant, à des fins de recherche, de comparer l'erreur quadratique moyenne engendrée par la règle 3 et celle produite par une règle selon laquelle on choisirait un commerce de remplacement dans la même zone de k commerces plutôt que laisser l'échantillon perdre une unité. Cette dernière solution est conditionnellement sans biais mais inconditionnellement biaisée.

Deux autres problèmes doivent être étudiés dans le cas d'une fermeture de commerce. Le premier concerne la coordination des mises à jour des ouvertures et des fermetures. L'ouverture et la fermeture des commerces se produit naturellement à des intervalles irréguliers, selon la situation économique et la croissance de la population. Pendant certaines périodes, il peut n'y avoir ni ouverture ni fermeture de commerces. Pendant d'autres périodes, certains commerces peuvent faire leur apparition sans qu'il n'y ait de fermeture ou vice versa. Tandis que dans d'autres périodes, il y aura à la fois des ouvertures et des fermetures. En théorie, il est possible d'utiliser différents cycles de mise à jour pour les ouvertures et les fermetures de commerces. Par exemple, la mise à jour peut être bimestrielle, mais alterner pour l'ouverture et la fermeture. Cette approche à l'avantage de régulariser la charge de travail. Cependant, les cycles alternatifs nuisent à la capacité de l'échantillon de bien mesurer les tendances, créant un effet en dent de scie dans la série chronologique des commerces puisque les ouvertures sont incluses dans l'échantillon avant que les fermetures en soient éliminées. Nous recommandons de faire les mises à jour au même moment afin de préserver les tendances.

Les probabilités de sélection inconditionnelles sont calculées ainsi:

$$\pi_{ijb} = k^{-1} p_{ijb}$$

pour $b = 1, \dots, B_{ij}$. Cela signifie que π_{ijb} est égal à la probabilité de choisir l'unité primaire d'échantillonnage, multipliée par la probabilité conditionnelle de sélection du commerce, étant donné l'U.P.E. sélectionnée.

Posons maintenant X'_{ijb} comme la valeur de l'unité U_{ijb} et X'_{ij+} comme le total de la (i,j) ième U.P.E. L'estimateur non biaisé du total de population X'_{ij} se calcule alors comme suit:

$$X'_{ij} = \sum_{n_i}^{j=1} Y'_{ijb} / \pi_{ijb},$$

où Y'_{ijb} est la valeur de l'unité tirée du (i,j) ième segment sélectionné, avec comme variance

$$\text{Var}\{X'_{ij}\} = \frac{1}{k} \sum_{i=1}^k \left(k \sum_{n_i}^{j=1} Y'_{ij+} - X'_{ij+} \right)^2 + k \sum_{n_i}^{j=1} \sum_{k}^{i=1} \sigma_{ij}^2, \quad (1)$$

où

$$\sigma_{ij}^2 = \sum_{B_{ij}}^{b=1} p_{ijb} \left(\frac{X'_{ijb}}{p_{ijb}} - X'_{ij+} \right)^2.$$

Le premier terme du membre de droite de l'équation (1) est la variance due à l'échantillonnage des segments. C'est la variance initiale puisque c'est la variance qui s'appliquait au moment du choix de l'échantillon original. Le second terme est la variance due au sous-échantillonnage à l'intérieur des segments. Noter que la valeur σ_{ij}^2 disparaît pour tout segment où il n'y a pas eu de sous-échantillonnage de nouveaux commerces. Noter aussi que la variance du sous-échantillonnage est minimale lorsque, pour chaque i et j donné, les probabilités p_{ijb} sont proportionnelles à X'_{ijb} . Dans ce cas, la composante de la variance à l'intérieur des segments s'annule. Dans une application réelle, toutefois, cette condition de proportionnalité n'est satisfait qu'en partie.

Comme d'habitude, une approximation de Taylor du premier ordre peut être utilisée pour connaître la variance de l'estimateur par quotient. Voir Wolter (1986) pour les techniques appropriées d'estimation de la variance de l'estimateur non biaisé X'_{ij} et de l'estimateur par quotient $X'_{R'ij}$.

À mesure que le temps passe, il faut mettre à jour l'échantillon pour refléter les ouvertures de commerces et les autres changements de la population. Il peut être bon de prévoir une mise à jour à des intervalles réguliers afin de faciliter le travail. Ces intervalles sont appelés cycles de mise à jour. De tels cycles peuvent être mensuels, bimestriels ou selon les exigences d'une application en particulier. Les facteurs à considérer dans l'établissement du cycle de mise à jour sont le coût de la mise à jour, la précision recherchée des estimateurs de niveau et de tendance et les besoins des clients ou des utilisateurs des données.

De façon générale, un échantillonnage mis à jour fréquemment sera plus coûteux mais plus précis et plus apprécié des clients qu'un échantillon mis à jour moins souvent. Pour un cycle de mise à jour à un temps donné t' , les règles 1 ou 1A peuvent être utilisées pour la mise à jour de l'échantillon. Les nouveaux commerces sont automatiquement placés dans le segment approprié à l'aide de leurs valeurs de Peano et l'indice b reflète cet ordre pour chaque cycle. Pour bien représenter ces notions, il aurait fallu ajouter la notion de temps en indice aux valeurs de U' 's, de B' 's, de p' 's, et de π' 's, mais nous ne l'avons pas fait pour faciliter la notation. Les formules des estimateurs de totaux X'_{ij} et $X'_{R'ij}$ et des variances correspon-

Définissons maintenant N segments de Peano, S_{ij} , en divisant l'intervalle $[P_L, P_U]$ en N tranches de P_{ij} . Posons $S_{ij} = [P_{ij}, P_{i+1,j})$, où $P_{k+1,j}$ représente $P_{i,j+1}$. Une définition spéciale est nécessaire pour le dernier segment. Posons $S_{knk} = [P_{knk}, P_U] \cup [P_L, P_1)$ afin que toute l'échelle de Peano $[P_L, P_U]$ soit couverte par les N segments. Cette définition spéciale, qui permet de définir l'échelle de Peano comme si elle était un cercle, est nécessaire pour s'assurer que la probabilité de sélection de tous les nouveaux commerces est différente de zéro. D'autres plans de segmentation peuvent être utilisés sans nuire aux propriétés statistiques du système de mise à jour. Notre système de mise à jour est basé sur les segments de Peano. Il permet de sélectionner les segments de façon systématique et de faire un sous-échantillonnage des commerces à l'intérieur des segments choisis. Ainsi, c'est le segment qui est l'unité primaire d'échantillonnage (U.P.E.) et non le commerce. Bien entendu, au moment du choix de l'échantillon initial, chaque segment ne compte qu'un seul commerce.

4.1 Échantillonnage des nouveaux commerces

Il est possible qu'un ou plusieurs commerces voient le jour à une période future t' du sondage. On assigne alors à chaque nouveau commerce une valeur de Peano unique qui fait partie d'un segment de Peano. La valeur de Peano permet de placer les nouveaux commerces automatiquement à l'endroit approprié dans la liste ordonnée de l'univers de l'enquête. La règle la plus simple pour l'échantillonnage des nouveaux commerces est la suivante:

Règle 1: Un nouveau commerce est inclus dans l'échantillon si et seulement si sa valeur de Peano fait partie d'un segment de Peano sélectionné. Les nouveaux commerces dont la valeur fait partie d'un segment non sélectionné ne sont pas inclus dans l'échantillon.

Selon cette règle, la probabilité de sélection d'un nouveau commerce est de $1/k$. Cela se produit parce que la probabilité de sélection du segment, qui est unique, est de $1/k$. Malheureusement, la règle 1 ne favorise pas un bon contrôle de la taille de l'échantillon dans le temps. Pour contrôler la taille de l'échantillon, nous proposons une forme de sous-échantillonnage dans les U.P.E. Supposons que $U_{i1}, U_{i2}, \dots, U_{iB_{ij}}$ soient les commerces du segment S_{ij} . Le commerce original est U_{i1} alors que les commerces $U_{i2}, U_{i3}, \dots, U_{iB_{ij}}$ sont les nouveaux commerces dans l'ordre de Peano. Le nombre de nouveaux commerces, $(B_{ij} - 1)$, dans un segment donné sera dans la plupart des cas 0, 1 ou 2. Ainsi, nous pouvons effectuer le sous-échantillonnage de la façon décrite ci-après.

Règle 1A. Un nouveau commerce pourra faire partie d'un sous-échantillon si et seulement si sa valeur de Peano fait partie d'un segment de Peano sélectionné. On associe les commerces $U_{i1}, U_{i2}, \dots, U_{iB_{ij}}$ aux probabilités $p_{i1}, p_{i2}, \dots, p_{iB_{ij}}$ où $p_{ijb} > 0$ et $\sum p_{ijb} = 1$. On choisit ensuite un des commerces selon cette mesure de probabilité. Le sous-échantillonnage est indépendant d'un segment choisi à l'autre. Les nouveaux commerces dont les valeurs de Peano font partie d'un segment non sélectionné ne sont pas inclus dans l'échantillon.

Les probabilités dans la règle 1A peuvent être égales ou inégales. Si elles sont inégales, elles peuvent être définies par rapport à une mesure provisoire de taille ou de manière à accélérer ou à retarder le remplissage de l'échantillon.

Les principaux objectifs de mise à jour sont remplis par la règle 1A. La règle permet de maintenir l'équilibre géographique dans le temps puisqu'il n'y a qu'une unité choisie à partir de chaque segment tiré à l'origine, segments qui sont eux-mêmes équilibrés géographiquement en raison du plan d'échantillonnage systématique. Deuxièmement, la règle maintient constante la taille de l'échantillon dans le temps puisqu'il n'y a toujours qu'un seul commerce choisi dans chacun des segments d'origine. Troisièmement, la règle est en accord avec les principes de l'échantillonnage probabiliste, où les probabilités de sélection doivent être connues et différentes de zéro; ainsi des estimateurs non biaisés du total de population peuvent être obtenus. Enfin, par un choix approprié de p_{ijb} , il est possible de contrôler la distorsion dans les tendances annuelles.

Dans l'application visée, les échantillonnages économiques sont placés sur une liste dans l'ordre de Peano à l'aide de leurs coordonnées de latitude et de longitude. Des échantillons aléatoires d'établissements peuvent être tirés systématiquement de cette liste. Comme les coordonnées géographiques de la terre sont des données stables, la position des nouveaux établissements sur la liste peut être établie sans équivoque. Ces établissements peuvent donc aussi être échantillonnés.

Pour illustrer cette application, voyons la figure 3 qui montre une chaîne de commerces de détail aux États-Unis. Chaque établissement est représenté par un code à deux lettres. L'ordre de Peano des établissements suit l'ordre alphabétique des codes.

Le chapitre suivant traite du système de mise à jour basé sur la liste des établissements en ordre de Peano.

4. RÈGLES DE MISE À JOUR D'UN ÉCHANTILLON

Dans les paragraphes suivants, nous décrivons un système de mise à jour des échantillons de commerces de détail. Comme nous l'avons dit plus tôt, ce système a été développé pour les applications de la A.C. Nielsen Company.

Supposons une strate d'échantillonnage donnée et arbitraire de taille N , où les commerces sont répartis selon l'ordre de Peano. Par exemple, une strate peut comprendre tous les commerces dans un marché métropolitain donné, comme Vancouver ou Montréal. La répartition par les valeurs de Peano est bien adaptée au système de mise à jour ci-après. D'autres types de répartition peuvent être utilisés, à condition qu'ils soient stables dans le temps et qu'ils puissent relier R^2 à R^1 de façon à conserver la contiguïté géographique et à donner à chaque nouveau commerce une position unique dans la liste ordonnée des commerces.

Supposons qu'un échantillon original de commerces est prélevé systématiquement avec probabilités égales dans la liste de Peano au temps $t = 0$. Soit U_{ij} , le commerce j dans l'échantillon systématique i possible, pour $i = 1, \dots, k$ et $j = 1, \dots, n_i$; k étant l'intervalle d'échantillonnage et n_i la taille de l'échantillon systématique i . Si $N = nk + r$, $r < k$, r échantillons seront de taille $n_i = n + 1$ et $k - r$ échantillons seront de taille $n_i = n$. Dans les paragraphes suivants, l'indice i est utilisé pour représenter l'échantillon choisi.

Soit P_{ij} la valeur de Peano associée à U_{ij} . Prenons P_L et P_U comme, respectivement, les plus petites et plus grandes valeurs de Peano possibles pour le marché à l'étude. Ainsi,

$$P_L \leq P_{11} < P_{21} < \dots < P_{k1} < P_{12} < \dots < P_{ij} < \dots < P_{knk} \leq P_U.$$

Nous supposons que chaque commerce n a qu'une seule situation géographique et donc une seule valeur de Peano.

Soit Y_{itj} la valeur d'une certaine caractéristique de U_{ij} au temps t . Un estimateur standard non biaisé de la population totale, Y_t , est le suivant:

$$Y_{it} = k \sum_{n_i}^{j=1} y_{itj},$$

tandis que l'estimateur par quotient est défini

$$Y_{Rit} = Y_{it} X_i / X_{it},$$

où la variable X est une mesure de la taille et X_i et X_{it} sont analogues à Y_i et Y_{it} respectivement.

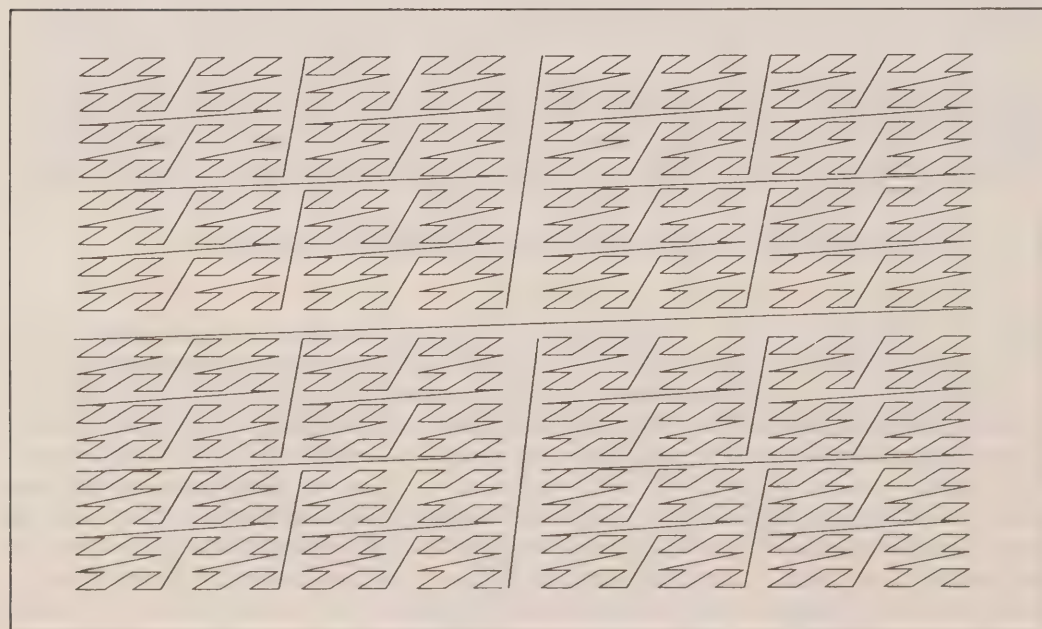


Figure 2. Courbe de Peano sur 1024 points

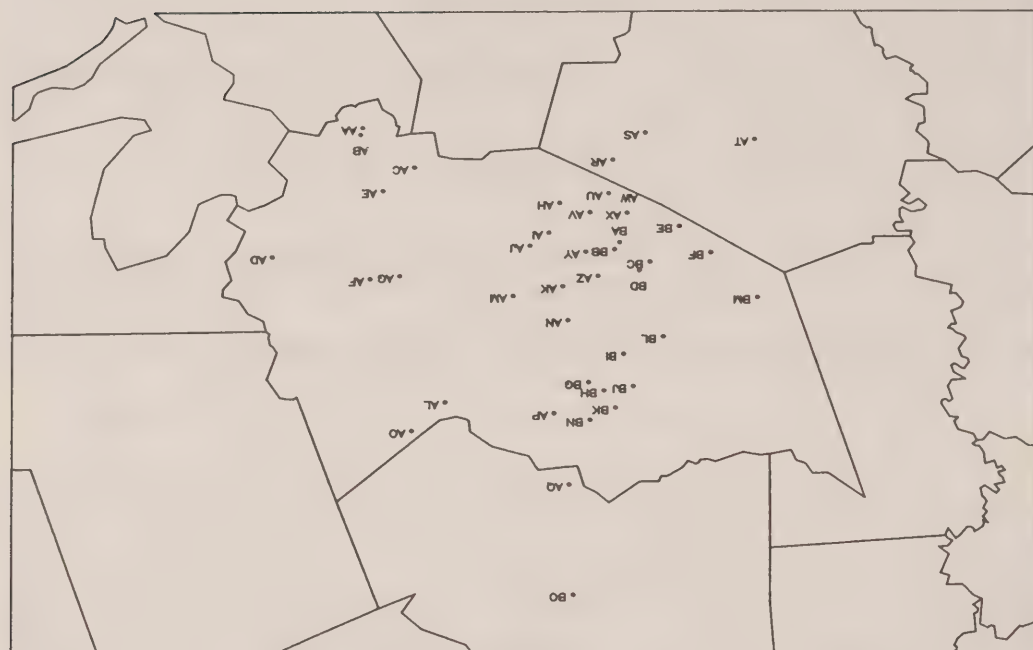


Figure 3. Chaîne de commerces de détail dans l'ordre de Peano

3. VALEURS DE PEANO

La valeur de Peano est un paramètre qui permet de définir une certaine courbe de remplissage fractale. Cette valeur établit une transformation de \mathbb{R}^2 dans \mathbb{R}^1 de sorte que les points dans \mathbb{R}^2 ou les objets dans l'espace sont arrangés dans un ordre unique (ordre de Peano) sur une liste. Dans la présente application, les objets sont les unités d'échantillonnage et l'espace \mathbb{R}^2 est représenté par le système de coordonnées géographiques de la terre.

La valeur de Peano est obtenue par entrecroisement de nombres binaires (voir Peano (1908), Laurini (1987) et Saalfeld, Fritfield, Broome et Meixler (1988)). Soit $X = X_k \dots X_3 X_2 X_1$ et $Y = Y_k \dots Y_3 Y_2 Y_1$ la longitude et la latitude d'un point arbitraire en format binaire de k chiffres. Alors la valeur de Peano correspondante sera $P = X_k Y_k \dots X_3 Y_3 X_2 Y_2 X_1 Y_1$. Voir la figure 1 pour un exemple dans le cas où $k = 4$. Noter comme il est simple de calculer la valeur de P .

Etant donné (pour tout k fini) des coordonnées de latitude et de longitude de k chiffres, le "point" spatial représenté par la valeur de P est en fait un carré dans \mathbb{R}^2 . A mesure que k augmente, la taille des carrés diminue. En fait, lorsque k tend vers l'infini, la valeur de P tend à représenter un point spécifique dans \mathbb{R}^2 .

La courbe de remplissage créée par les valeurs de Peano prend la forme d'un N récursif. La courbe en N est illustrée à la figure 2 sur une grille de 1024 points. Cette figure montre bien l'aspect récursif des images fractales.

La courbe en N passe une fois et une fois seulement par chaque point de l'espace, les points étant des carrés dont la taille est déterminée par le nombre de chiffres des coordonnées de longitude et de latitude. L'ordre des points sur la courbe (ordre de Peano) préserve largement la contiguïté géographique, ce qui facilite les recherches de proximités. L'ordre de Peano entraîne quelques discontinuités géographiques (saut du point 516 au point 517 dans la figure 2), comme toute correspondance entre \mathbb{R}^2 et \mathbb{R}^1 .

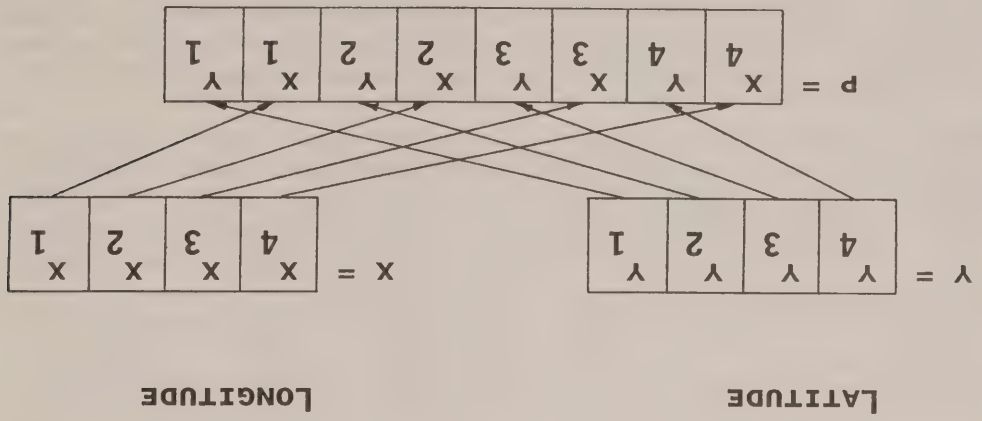


Figure 1. Calcul de la valeur de Peano par entrecroisement de nombres binaires

L'échantillon Scantrack doit être continuellement mis à jour parce que l'univers de l'enquête est composé d'environ 30,500 supermarchés n'est pas statique. Lors d'une période récente de 12 mois, environ 2,200 nouveaux supermarchés ont vu le jour et 2,450 ont fermé leurs portes. De plus, 170 établissements ont été reclassifiés. La reclassification peut être causée par différents facteurs. Certains petits magasins d'alimentation font leur entrée dans l'univers Scantrack lorsque leur volume de ventes devient supérieur à 2 millions de dollars par année, ce qui les fait entrer dans la catégorie des supermarchés. Un magasin peut déménager, changer de nom ou être agrandi. Certains supermarchés peuvent changer de statut et devenir des hypermarchés, des magasins-entrepôts ou un autre type de supermarché non traditionnel. En 1979, les quelque 3,800 hypermarchés et magasins-entrepôts étaient responsables de 17% des ventes totales des supermarchés. En 1988, leur nombre est passé à 9,000 et ils étaient responsables de près de 50% des ventes totales (*Progressive Grocer* 1989). Dans certains cas, des supermarchés ou même des chaînes entières de supermarchés sont acquis par une autre organisation, ce qui influe sur la définition des strates.

En plus des changements de population, le manque de données ou leur inexactitude peut entraîner le remplacement de certaines unités d'échantillonnage. Certains supermarchés sélectionnés ne possèdent pas de lecteurs optiques ou ont des lecteurs optiques qui ne sont pas compatibles. Si un supermarché ne peut fournir les données utiles de façon constante, il doit être éliminé de l'échantillon. Dans certains cas, la demande de modification de l'échantillon provient de la chaîne elle-même. Occasionnellement, un détaillant peut simplement refuser de coopérer.

Les principaux objectifs du système de mise à jour pour l'échantillon Scantrack sont les suivants:

- 1) maintenir l'équilibre géographique de l'échantillon dans le temps,
- 2) maintenir la taille de l'échantillon dans le temps,
- 3) s'assurer que les principes de l'échantillonnage probabiliste sont respectés afin d'éviter que les estimateurs des ventes totales ne soient biaisés et
- 4) s'assurer que les changements dans l'échantillon ne modifient pas excessivement les estimations des tendances annuelles.

L'équilibre géographique constitue une approximation pour l'équilibre socio-économique. Comme différents quartiers ont différentes habitudes de consommation, l'équilibre géographique est important pour que le plan d'échantillonnage soit efficace (*i.e.*: faible variabilité d'échantillonnage) pour un grand nombre d'articles. De plus, les clients considèrent que l'équilibre géographique est une condition importante pour un échantillonnage adéquat.

La diminution de la taille d'un échantillon fait augmenter l'erreur-type des estimateurs tandis que son augmentation accroît le coût du sondage. Les deux situations sont à éviter. De plus, les contrats passés avec les organisations indiquent la taille des échantillons et les paiements, et tout changement doit être renégocié. Cette situation est aussi à éviter.

Toutes les applications utilisant les données Scantrack requièrent des estimateurs efficaces et non biaisés des ventes totales. Les fabricants et les commerçants utilisent ces données couramment pour prendre des décisions: quelle quantité de produits faut-il fabriquer, quelle quantité faut-il distribuer, quelle quantité garder en stock, comment organiser les étalages, etc.

Les clients ont également besoin d'estimations fiables sur les tendances annuelles pour la gestion de leurs affaires. Ces données permettent aux fabricants d'évaluer la santé de leur entreprise. Il est profitable, à la fois pour le commerçant et pour le fabricant, de connaître le rendement à long terme de toutes les marques importantes dans toutes les catégories de produits.

Le système de mise à jour élaboré pour remplir ces objectifs est décrit au chapitre 4. Mais d'abord, un nouveau plan de répartition géographique est décrit au chapitre 3.

principaux sondages, le sondage Scantrack, et les problèmes de mise à jour auxquels nous sommes confrontés dans ce sondage. Nous décrivons aussi certains des principaux objectifs que nous poursuivons en élaborant un nouveau système de mise à jour pour ce sondage. Le nouveau système de mise à jour est basé sur un paramètre connu en mathématique comme les valeurs de Peano, qui permettent de créer une courbe de remplissage fractale. Les valeurs de Peano sont définies dans le chapitre 3, qui comprend aussi plusieurs graphiques pour faciliter la compréhension. La conclusion est donnée au chapitre 4, qui comprend la description des règles du nouveau système de mise à jour de l'échantillon.

2. SONDAGE SCANTACK

Les sociétés Nielsen fournissent des données provenant de plusieurs études de marché. L'unité d'échantillonnage des sondages portant sur les médias, comme le Nielsen Television Index et le Nielsen Station Index, est soit le logement ou le ménage. Pour les sondages portant sur l'industrie des biens de consommation comme le Nielsen Food Index, le Nielsen Drug Index et le Nielsen Scantrack United States (NSUS), ce sont les magasins qui sont les unités d'échantillonnage. Le Single Source Service, qui étudie les habitudes de consommation en fonction de l'écoute de la télévision et de la commercialisation des produits, se sert des ménages et des magasins comme unités d'échantillonnage. Bien que la mise à jour de l'échantillon soit importante pour chacune de ces enquêtes, la présente discussion porte sur l'échantillon Scantrack de supermarchés, qui est à la base du service NSUS. L'échantillon Scantrack compte 3,000 supermarchés répartis en 51 strates, 50 pour les grandes villes et une pour le reste des Etats-Unis. A l'intérieur d'une même strate ou marché, l'échantillon est stratifié à nouveau en fonction des grandes chaînes de supermarchés. La base est ordonnée géographiquement et un échantillon systématique est choisi à l'intérieur de chaque strate afin de bien représenter la situation socio-économique. Cet échantillon est aussi représentatif de l'âge du supermarché, de sa taille et des autres facteurs qui influent sur les ventes. Bien qu'un échantillon systématique ordonné géographiquement soit très simple et direct, le choix de ce plan d'échantillonnage est justifié par des années d'expérience et les résultats d'études empiriques portant sur l'essai de divers plans d'échantillonnage dans des conditions réelles.

Les supermarchés qui composent l'échantillon Scantrack sont munis de lecteurs optiques qui lisent les codes à barres sur les emballages au moment de l'achat. Les codes à barres sont appelés codes universels de produits ou CUP. Lorsque l'étiquette du produit est lue, la transaction est enregistrée dans l'ordinateur du magasin et le prix correspondant au CUP est enregistré. Chaque semaine, le commerce nous indique le total des ventes et le prix de chaque article vendu. Comme le nombre d'articles portant un code CUP peut aller jusqu'à 10,000 dans un seul supermarché, nous traitons plus de 30 millions d'observations par semaine.

En plus des données des lecteurs optiques, nous recueillons les données portant sur la commercialisation d'un produit, qu'il s'agisse des coupons-rabais, de la publicité parue dans un journal ou de la disposition du produit dans le magasin. Lorsque un produit a fait l'objet d'une publicité, le type de publicité imprimée utilisée et l'emplacement de l'étalage dans le magasin sont aussi connus.

Les rapports NSUS comprennent le total estimé des ventes de chaque article et groupe d'articles pour chacun des marchés et pour l'ensemble des Etats-Unis. Un estimateur par le quotient est utilisé et la variable auxiliaire est le volume global des ventes. Le volume global des ventes est le total des ventes de tous les articles d'un magasin, habituellement sur une période d'un an. Cette valeur est en étroite corrélation avec les ventes et des taux de vente en fonction du rapport NSUS comprend des estimations des ventes et des taux de vente en fonction des conditions de commercialisation des produits, ainsi que des estimations des tendances annuelles de vente.

Pour que l'échantillon reflète bien les changements qui se produisent dans l'univers de l'enquête et par le fait même demeure représentatif, l'organisation responsable du sondage doit élaborer un système de mise à jour explicite. Un *système de mise à jour* comprend un plan d'échantillonnage et une méthode de mise à jour de la base de sondage, vraisemblablement énoncés sous forme de règles simples. Ces règles permettent au statisticien de bâtir l'échantillon de façon que la probabilité que chaque unité élémentaire soit incluse dans l'échantillon soit connue et différente de zéro pour toutes les périodes du sondage successif ou, si cela est impossible, de pondérer les données du sondage correctement afin d'en arriver à des estimateurs non biaisés ou convergents des paramètres d'intérêt. Dans les cas i) à iv), il est évident que le système de mise à jour doit remplir au moins les quatre conditions suivantes:

- Donner aux nouvelles unités élémentaires une probabilité de sélection connue et différente de zéro.
- Tenir compte des unités élémentaires qui n'existent plus réellement.
- Empêcher que les unités élémentaires soient incluses plusieurs fois dans l'échantillon; si c'est impossible, le système doit tenir compte de la situation afin que les ajustements nécessaires puissent être faits au cours de l'estimation.
- Mettre à jour la base de sondage afin de faciliter et de contrôler les fonctions ci-dessus.

Tous les systèmes de mise à jour doivent respecter la règle suivante: le système ou les règles qui le définissent doivent traiter de façon symétrique les changements de l'univers de l'enquête, qu'ils se produisent à l'intérieur ou à l'extérieur de l'échantillon. Si un système de mise à jour ne respecte pas cette règle, les estimateurs de totaux et des autres paramètres de la population risquent d'être biaisés. Par exemple, considérons deux règles qui peuvent être utilisées dans le cas ii) pour l'échantillonnage de nouvelles sociétés créées par transfert. Une des possibilités est d'inclure les nouvelles sociétés dans l'échantillon si les sociétés mères faisaient partie de l'échantillon. Si ce n'est pas le cas, la nouvelle société peut faire l'objet d'un nouvel échantillonnage. Ainsi, la probabilité de sélection des nouvelles sociétés est multiple et peut conduire à une estimation biaisée, à moins que des ajustements soient effectués (ces ajustements se rapportent aux règles de multiplicité étudiées par Monroe Sirken (1970) et d'autres). La seconde possibilité est d'inclure les nouvelles sociétés dans l'échantillon si les sociétés mères faisaient partie de l'échantillon. Comme cette seconde règle traite de façon symétrique les changements dans l'univers du sondage qui se produisent dans l'échantillon et à l'extérieur de celui-ci, l'estimation du paramètre d'intérêt n'est pas biaisée.

Pendant l'élaboration d'un système de mise à jour, le statisticien ne doit pas se baser uniquement sur les propriétés statistiques des estimateurs, mais aussi sur le coût et l'applicabilité des règles pertinentes; il doit aussi s'assurer que ces règles seront perçues favorablement par le client. Certaines règles peuvent exiger une collecte de données supplémentaires, entraînant par le fait même des coûts additionnels qui doivent être prévus dès la mise sur pied d'un nouveau sondage successif. Dans certains cas, les données additionnelles peuvent devoir être recueillies après la fin de l'étude de façon rétrospective. Cela peut s'avérer difficile, ou à tout le moins, peut entraîner des erreurs considérables non dues à l'échantillonnage, ce qui peut introduire un biais. Certaines règles peuvent être applicables et peu coûteuses, sans toutefois satisfaire aux exigences du client ou des utilisateurs des données.

Le problème de la mise à jour des échantillons n'est pas un problème nouveau; des systèmes de mise à jour des échantillons sont utilisés depuis nombre d'années dans plusieurs des enquêtes permanentes de Statistique Canada, du United States Bureau of the Census et de la A.C. Nielsen Company. Néanmoins, il existe très peu d'ouvrages sur le sujet. Pour une brève discussion des problèmes de mise à jour, voir Wolter et coll. (1976) pour le cas ii), Hanson (1978) pour le cas iii) et Ernst (1989) pour le cas iv). Lire également les commentaires généraux de Duncan et Kalton (1987) sur les enquêtes-ménages et ceux de Colledge (1989) sur les enquêtes-entreprises.

Dans la suite du présent article, nous étudierons principalement le cas i), où l'établissement est à la fois l'unité d'échantillonnage et l'unité élémentaire. C'est ce qui se produit dans nos sondages sur les établissements à la A.C. Nielsen Company. Le chapitre 2 décrit un de nos

légale propriétaire du commerce, ou la commission scolaire et ainsi de suite. Dans certains cas, l'établissement et la société sont une seule et même chose, par exemple, dans le cas d'un magasin d'alimentation indépendant.

Dans le cas i), l'univers de l'enquête peut varier comme suit:

- Etablissement nouvellement construit.
- Etablissement qui passe d'une catégorie non observée à une catégorie observée.
- Etablissement qui passe d'une catégorie observée à une autre catégorie observée.
- Etablissement qui passe d'une catégorie observée à une catégorie non observée.
- Changement de fonction d'un édifice, d'une fonction résidentielle à une fonction commerciale.
- Changement de fonction d'un édifice, d'une fonction commerciale à une fonction résidentielle.
- Démolition d'un établissement existant.
- Etablissement occupé de façon irrégulière.
- Changement de configuration d'un établissement, par exemple, s'il est divisé en deux ou plusieurs établissements.

Le cas ii) est beaucoup plus complexe que le cas i) puisque les unités d'échantillonnage sont des grappes d'unités élémentaires. Toutes les situations qui se produisent dans le cas i) peuvent aussi se produire dans le cas des sociétés constituées d'un seul établissement. Pour les sociétés formées de plusieurs établissements, les problèmes suivants s'ajoutent:

- Fusion de deux sociétés pour constituer une nouvelle société.
- Absorption d'une société par une autre. La société absorbante est la seule société résultante.
- Entreprise conjointe, où deux sociétés collaborent pour former une nouvelle société qui peut être une filiale des deux sociétés mères.
- Opération de transfert où une société crée une nouvelle société indépendante.
- Opération de transfert où une société vend une partie de ses parts à une autre société.

Les situations qui se produisent dans le cas iii) ressemblent beaucoup à celles qui se produisent dans le cas i):

- Habitation nouvellement construite.
- Habitation qui passe d'une catégorie non observée à une catégorie observée.
- Habitation qui passe d'une catégorie observée à une autre catégorie observée.
- Habitation qui passe d'une catégorie observée à une catégorie non observée.
- Changement de fonction, de résidentielle à commerciale.
- Changement de fonction, de commerciale à résidentielle.
- Démolition d'habitations existantes.
- Reconfiguration d'édifices existants, par exemple, reconfiguration des appartements dans un petit édifice à plusieurs logements.

Enfin, le cas iv) ressemble beaucoup au cas ii) en ce qui a trait à la composition de la population de l'enquête et à la complexité des changements qui s'y produisent. Les causes de mise à jour sont les suivantes:

- Mariage, création d'une nouvelle famille à partir de familles existantes ou de parties de familles.
- Nouveaux membres qui entrent dans une famille existante, ce qui élimine une autre famille ou partie de famille.
- Divorces, qui peuvent mener à la création d'une nouvelle famille à partir d'une famille existante.
- Déménagement des membres d'une famille, soit pour joindre une autre famille ou en créer une nouvelle.
- Naissances.
- Décès.
- Déménagement d'une famille entière; il faut alors les retrouver et peut-être modifier les charges de travail assignées.

Mise à jour des échantillons basée sur les valeurs de Peano

KIRK M. WOLTER et RACHEL M. HARTER¹

RÉSUMÉ

Le présent document porte sur la mise à jour des échantillons et des bases utilisées dans les enquêtes à passages répétés. Le système de mise à jour décrit remplit quatre objectifs principaux: 1) maintenir une répartition géographique équilibrée de l'échantillon, 2) maintenir constante la taille de l'échantillon, 3) maintenir le caractère non biaisé de l'estimateur et 4) empêcher l'apparition de distorsion dans l'estimation des tendances. Le système est basé sur les valeurs de Peano qui permettent de créer une courbe de remplissage fractale. L'exemple utilisé pour présenter le nouveau système est un sondage à l'échelle nationale portant sur les établissements des États-Unis, sondage effectué par la A.C. Nielsen Company. MOTS CLÉS: Enquêtes à passages répétés; mise à jour d'échantillon; unités de population changeantes; valeur de Peano.

1. INTRODUCTION

Dans le présent document, nous étudierons les enquêtes à passages répétés et la mise à jour qu'elles exigent. Soit U_t l'univers d'un sondage au temps t , $t = 0$ signifiant le début d'un nouveau sondage. Supposons aussi qu'un échantillon aléatoire d'unités de U_0 a été choisi et qu'il est donc possible de construire des estimateurs non biaisés (ou du moins convergents) du total pour la population et des autres paramètres d'intérêt. Cet univers est sondé régulièrement dans le temps, afin de déterminer le "niveau" de la population et de mesurer ses tendances. Dans de telles enquêtes portant sur des personnes ou des institutions, les deux seuls points d'intérêt dans notre cas, la composition de la population varie avec le temps, en raison des naissances, des décès et des autres changements qui se produisent dans les unités d'échantillonage. La base de sondage, le plan d'échantillonage et les méthodes d'observation ou de collecte des données doivent être constamment révisés en fonction de tels changements; autrement, l'échantillon devient très biaisé et n'est plus représentatif de l'univers de l'enquête. Le type de problèmes de mise à jour qui se produisent dans les sondages successifs dépend à la fois de l'univers étudié, du choix de l'unité d'échantillonage et des interactions entre l'unité d'échantillonage et les unités élémentaires de l'univers de référence. Nous donnerons un bref résumé des problèmes qui se produisent dans quatre cas différents:

- i) enquête portant sur les établissements; unité d'échantillonage: l'établissement.
- ii) enquête portant sur les établissements; unité d'échantillonage: société ou autre regroupement similaire d'établissements.
- iii) enquête portant sur des personnes ou des ménages; unité d'échantillonage: adresse ou logement.
- iv) enquête portant sur des personnes ou des ménages; unité d'échantillonage: le ménage ou la famille.

Dans cet article, les mots "établissement" et "société" sont utilisés dans leur sens général. Un établissement peut être un commerce de détail, une usine de fabrication, une école, un hôpital, un terrain de golf ou toute autre entité localisée, tandis que la société désigne l'entité

¹ Kirk M. Wolter et Rachel M. Harter, Statistical Research Department, A.C. Nielsen Company, Nielsen Plaza, Northbrook, IL 60062, E.-U.

- ECKLER, A.R. (1955). Rotation sampling. *The Annals of Mathematical Statistics*, 26, 664-685.
- FULLER, W.A. (1986). Small area estimation as a measurement error problem. *Proceedings of the Conference on Survey Research Methods in Agriculture*, (éd. D. Faulkenberry), American Statistical Association and NASS, U.S. Department of Agriculture, Washington, D.C.
- GARCIA, P.A., BATTESE, G.E., et BREWER, W.D. (1975). Longitudinal study of age and cohort influences on dietary patterns. *Journal of Gerontology*, 30, 349-356.
- GRAHAM, J.E. (1973). Composite estimation in two cycle rotation sampling designs. *Communications in Statistics*, 1, 419-431.
- GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics, American Statistical Association*, 242-257.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 300-303.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, séries B, 42, 221-226.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. University of Michigan, Survey Research Center.
- KASPRZYK, D., DUNCAN, G.J., KALTON, G., et SINGH, M.P. (1989). *Panel Surveys*. New York: John Wiley.
- KASPRZYK, D., et McMILLLEN, D.B. (1987). SIPP: Characteristics of the 1984 Panel. *Proceedings of the Section on Social Statistics, American Statistical Association*, 181-186.
- LAZARSFELD, P.F., et FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.
- LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. Dans *Panel Surveys* (eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- LITTLE, R.J.A., et SU, H.L. (1989). Item Nonresponse. Dans *Panel Surveys* (eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley.
- MADOW, W.G., OLKIN, I., NISSELSOHN, H., et RUBIN, D.B. (1983). *Incomplete Data in Sample Surveys*. (Trois volumes) New York: Academic Press.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, séries B, 12, 241-255.
- POTERBA, J.M., et SUMMERS, L.H. (1985). Adjusting the gross change data: Implications for Labor Market Dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census et U.S. Bureau of Labor Statistics, 81-95.
- RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SINGH, A.C., et RAO, J.N.K. (1990). Adjustments for classification error in gross flows. Document non publié, Statistique Canada, Ottawa, Canada.
- SMITH, T.M.F., et HOLT, D. (1989). Some inferential problems in the analysis of surveys over time. Communication présentée à la 47^{ème} session de l'Institut International de Statistique, Paris.
- WOLTER, K. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Pour de petits échantillons, l'erreur quadratique moyenne est moins élevée dans le cas de l'estimateur direct parce que la variance de cet estimateur est moins élevée. Rappelons-nous que seulement les trois quarts des observations nous renseignent sur $P^{EE} = P_{11}$. Toutefois, pour des échantillons dont la taille est supérieure à 750, le carré du biais de l'estimateur direct représente la très grande partie de l'erreur quadratique moyenne de cet estimateur, laquelle est plus élevée que celle de l'estimateur redressé en fonction de l'erreur de mesure. Ce court exemple illustre bien l'efficacité des plans de sondage qui permettent d'estimer les paramètres du processus d'erreur de mesure.

6. RÉSUMÉ ET CONCLUSIONS

Dans cet article, nous avons fait un survol des questions qui ont trait à l'analyse des données d'enquêtes à passages répétés. Nous avons vu qu'il est possible d'accroître sensiblement l'efficacité de l'aide de méthodes fondées sur les moindres carrés. Cependant, la complexité des échantillons et les délais fixés font qu'on ne peut tenir compte de toutes les données disponibles dans la construction des estimateurs par les moindres carrés. Le statisticien doit donc, en pratique, se limiter à un sous-ensemble de variables pour la construction des poids pertinents. En deuxième lieu, nous avons étudié la méthode d'estimation utilisée dans une enquête à deux passages réalisée par le U.S. Soil Conservation Service.

Nous avons montré que l'erreur de mesure pouvait introduire un biais appréciable dans des estimations longitudinales comme les estimations de variations brutes. Nous avons vu qu'il était possible de construire des estimateurs convergents à l'aide de méthodes fondées sur l'erreur de mesure. En outre, il est justifié de consacrer le quart des ressources disponibles à l'estimation de la variance de l'erreur de mesure dans le but explicite d'utiliser des méthodes d'estimation fondées sur la mesure de ce genre d'erreur.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à un contrat de coopération (n° 68-3A75-8-12) passé avec le Soil Conservation Service du Département de l'agriculture des E.-U. L'auteur remercie Margot Tollefson pour les calculs.

BIBLIOGRAPHIE

- ABOWD, J.M., et ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.
- BATTESE, G.E., et FULLER, W.A. (1973). An unbiased response model for analysis of categorical data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 202-207.
- BATTESE, G.E., HASABELNABY, N.A., et FULLER, W.A. (1989). Estimation des cheptels à l'aide de plusieurs estimateurs à base de sondage aréolaire et à base de sondage multiple. *Techniques d'enquête*, 15, 15-29.
- CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

du tableau 3. Les trois autres représentent les estimateurs redressés en fonction du biais dû à l'erreur de mesure. Dans le calcul de la variance, on suppose que α a une erreur type de 0.01. La correction effectuée pour tenir compte du biais dû à l'erreur de mesure ne modifie pas les estimateurs de P_E et P_{EE} . Dans cet exemple, le carré du biais de l'estimateur ordinaire de P_{EE} équivaut à environ neuf fois la variance de l'estimateur par les moindres carrés généralisés. Le biais dû à l'erreur de mesure explique donc en très grande partie l'erreur quadratique moyenne de l'estimateur de P_{EE} .

Ces résultats ont des conséquences importantes pour ce qui a trait à la conception du plan de sondage. Afin d'illustrer cela, revenons au problème de la variation brute. Supposons que nous voulions estimer la probabilité qu'une personne appartienne à la catégorie des personnes avec emploi pendant deux périodes données, P_{EE} . Nous supposons qu'il est possible de réaliser des réinterviews indépendantes aux deux périodes et que les interviews réalisées à deux périodes quelconques sont indépendantes. Nous supposons aussi qu'il n'y a que deux scénarios d'interview possibles:

A. Interview et réinterview à l'une des deux périodes.
B. Interview à la période 1 et interview à la période 2.

Enfin, nous supposons que l'erreur de réponse est non biaisée et qu'un modèle à deux catégories (personnes avec emploi et chômeurs) est approprié dans les circonstances. Nous supposons aussi que la probabilité qu'une réponse soit exacte dépend uniquement de la catégorie à laquelle appartient le répondant dans la période courante. Soient les probabilités de réponse définies en fonction de α et soit

$$\gamma = (1 - \alpha)^{-2}.$$

Désignons par θ_{ij} l'élément ij de la matrice de probabilités 2×2 observée par suite de la réinterview. Par conséquent, θ_{ij} est la probabilité qu'une personne réponde i à l'interview et j à la réinterview. Pour ce modèle simple, il existe des formules explicites pour les estimateurs. Ainsi,

$$\hat{\gamma} = (\hat{\theta}_{11} - \hat{\theta}_1^2)^{-1}(\hat{\theta}_{11} - \hat{\theta}_2^2)$$

et

$$\hat{P}_{11} = \hat{\gamma}(\hat{P}_{11} - \hat{P}_1.\hat{P}_1) + \hat{P}_1.\hat{P}_1$$

ou

$$\theta_1 = \theta_{11} + \theta_{21},$$

θ_{ij} sont les estimations découlant de la réinterview et \hat{P}_{ij} , les estimations découlant des interviews réalisées aux deux périodes.

Dans la construction de l'estimateur, les résultats de la réinterview ne servent qu'à estimer le paramètre de l'erreur de mesure. En fait, ces résultats pourraient être utilisés dans une méthode des moindres carrés généralisés pour améliorer les valeurs estimées de P_{11} , P_1 et P_1 . En supposant qu'il en coûte la même chose pour chaque interview, nous pouvons montrer qu'environ le quart des ressources devraient être consacrées à la réinterview. Le tableau 6 indique l'efficacité des estimateurs redressés en fonction de l'erreur de mesure par rapport aux estimateurs directs biaisés.

Tableau 6

Efficacité relative des estimateurs redressés en fonction de l'erreur de mesure

Taille de l'échantillon (n)		EQM (direct)/EQM (erreur de mesure)	
500	1,000	0.87	1.13
5,000	5,000	3.22	3.22
10,000		5.84	

Pour montrer jusqu'à quel point l'erreur de mesure peut biaiser les estimateurs de la variation brute, nous allons reprendre l'exemple du tableau 1. Chua et Fuller (1987) montrent que les proportions estimées qui figurent dans les cases du tableau à double entrée seront fortement biaisées si les données sont recueillies au moyen d'une méthode comme celle qu'utilise le U.S. Bureau of the Census. Voir aussi à ce sujet Abowd et Zellner (1985), Poterba et Summers (1985) et Singh et Rao (1990). Le modèle de Chua-Fuller suppose que les erreurs de réponse aux deux périodes sont indépendantes. Il suppose aussi que, pour chaque période,

$$\begin{aligned} P\{\text{réponse} = E|\text{situation réelle} = E\} &= 1 - \alpha + \alpha P_E, \\ P\{\text{réponse} = U|\text{situation réelle} = E\} &= \alpha P_U, \\ P\{\text{réponse} = U|\text{situation réelle} = U\} &= 1 - \alpha + \alpha P_U, \\ P\{\text{réponse} = E|\text{situation réelle} = U\} &= \alpha P_E, \end{aligned}$$

où α est le paramètre du mécanisme de réponse. Selon ce modèle, l'espérance de la proportion d'employés à n importe quelle période est égale à la proportion réelle. Un estimateur convergent de P_{EE} selon le modèle de Chua-Fuller est

$$\hat{\pi}_{EE} = (1 - \alpha)^{-2} \{ P_{EE} - P_E \cdot P_E [1 - (1 - \alpha)^2] \},$$

où P_{EE} , P_E , et P_U sont les estimateurs directs et α est un paramètre du mécanisme de réponse. Voir aussi Battese et Fuller (1973). Compte tenu des résultats de la réinterview, une valeur $\alpha = 0.10$ n'est pas exagérée. Par conséquent, nous avons

$$\pi_{EE} = (0.90)^{-2} \{ 0.91 - 0.93(0.94)(0.19) \} = 0.9184.$$

Le tableau des proportions corrigées en fonction de l'erreur de réponse est

$$\begin{pmatrix} 0.9184 & 0.0116 \\ 0.0216 & 0.0484 \end{pmatrix}.$$

Dans cet exemple, le biais de l'estimateur direct de P_{EE} est 0.0084. Chua et Fuller estiment le même biais à 0.0168 dans le tableau de contingence qui comprend aussi les inactifs. Dans le tableau 5, nous comparons des méthodes d'estimation pour P_{EE} . Nous supposons un échantillon de 10,000 unités. Les trois méthodes de la portion gauche du tableau sont celles

Tableau 5

Erreur quadratique moyenne de divers estimateurs pour un échantillon de 10,000 unités à chaque période et un taux de chevauchement de 50% (erreur quadratique moyenne de l'estimateur par les MCG redressé en fonction de l'erreur de mesure = 100)

Paramètre	Méthode d'estimation					
	Ordinaire		MCG	Erreur de mesure		
	Simple	MCG avec contrainte		Simple	MCG avec contrainte	
P_E	111	111	100	111	111	100
$P_{\cdot E}$	111	101	100	101	101	100
P_{EE}	1071	967	961	250	106	100

Tableau 4

Illustration de la méthode d'estimation

1982	1987				TOTAL
	Terrain labourable	Autres	Terrain urbain	Routes	
Terrain labourable	26,243	179	13	6	26,441
Autres	771	7,114	6	2	7,893
Terrain urbain	0	0	623	0	623
Routes	17	4	0	1,038	1,059
TOTAL pour 1987	27,031	7,297	642	1,046	36,016

importants comme le terrain urbain, le terrain labourable et les petites nappes d'eau. On s'est donc servi d'une méthode d'estimation pour petites régions pour établir des estimations de variation de superficie pour les zones principales de sols comprises dans les limites d'un comté (MLRAC). À cette fin, on a utilisé un programme d'ordinateur élaboré à l'Université Iowa State, Fuller (1986) expose la théorie qui est à la base de la méthode d'estimation pour petites régions. Le programme d'ordinateur a permis d'établir des estimations de la variation de superficie pour cinq modes d'utilisation du sol secondaires pour chacune des 5,500 MLRAC. Il s'agit là essentiellement d'une opération de répartition en ce sens que la somme des estimations pour les MLRAC équivalent à l'estimation globale pour l'Etat. Par ailleurs, on a établi des estimations pour les éléments du tableau 4 (élargi à 14 modes d'utilisation du sol) pour chaque MLRAC.

À cette occasion, les estimations régionales pour les MLRAC, les estimations de la superficie représentée par les routes et les estimations de la superficie totale de terrain labourable pour l'Etat ont servi de totaux de contrôle. Le processus d'estimation s'est terminé par la pondération des données des points d'observation de manière à obtenir les estimations du tableau 4 pour chaque MLRAC.

En résumé, le processus d'estimation aboutit à une série de données de totalisation qui se rapportent à des points d'observation et qui permettent d'estimer tous les éléments d'un tableau à double entrée décrivant, pour n'importe quelle région identifiable, l'évolution de la superficie associée à divers modes d'utilisation du sol pour la période 1982-1987. Les estimations ainsi obtenues concordent avec les estimations établies antérieurement à l'échelle de l'Etat pour les principaux modes d'utilisation du sol et concordent aussi avec les données provenant de sources autres que les points d'observation.

En règle générale, l'échantillon de totalisation ne produit pas de bonnes estimations de la variance même si les segments et les strates sont bien identifiés dans la série de données. À cause du contrôle exercé sur l'échantillon de 1982, qui était plus grand, les données des points d'observation relatives aux principaux modes d'utilisation du sol, comme le terrain labourable, produiront des estimations de la variance trop élevées. Pour obtenir des estimations justes, il faudrait recourir aux formules de l'échantillonnage double.

5. ERREUR DE MESURE

L'erreur de mesure peut avoir une incidence notable sur l'analyse des données dans le temps. Cette incidence peut être modérée dans le cas de moyennes observées périodiquement mais peut aussi être très appréciable dans le cas de l'estimation de la variation brute ou de l'estimation par régression.

pour tout le segment en ce qui concerne des aspects comme le sol urbain et les plans d'eau. En revanche, des données détaillées sur la nature et l'utilisation du sol sont recueillies à certains endroits dans le segment, choisis aléatoirement. En règle générale, on compte trois points d'observation par segment; les segments de 40 acres n'en comptent que deux et les segments des échantillons de deux États n'en comptent qu'un. Certaines données, comme la superficie totale et la superficie représentée par les routes, sont recueillies au moyen d'un recensement qui n'a rien à voir avec l'enquête précitée.

En 1982, l'échantillon comprenait environ 350,000 segments et près d'un million de points d'observation. En 1987, il comptait environ 100,000 segments, dont la majeure partie provenait de l'échantillon de 1982. Néanmoins, environ 1,500 nouveaux segments, prélevés dans des régions à forte croissance urbaine, ont été inclus dans l'échantillon de 1987. Ainsi, celui-ci comptait environ 280,000 points d'observation.

L'enquête de 1987 a été la première où on a décidé de faire une analyse de données longitudinales; cette analyse allait porter sur la période 1982-1987. Par la même occasion, on a décidé que les données de l'enquête allaient désormais être mises à la disposition du Soil Conservation Service de chaque État pour qu'il puisse faire ses propres analyses.

En 1987, les membres du personnel sur le terrain se sont vu remettre une feuille de travail qui contenait les données des segments pour 1982. Ils devaient y inscrire les données pour 1987 en se fondant sur les résultats d'observations sur le terrain et de la photographie aérienne. Ils étaient autorisés à corriger les données de 1982 si celles-ci étaient inexactes. Des méthodes de contrôle et de vérification ont été appliquées durant la phase de traitement.

On a conçu l'échantillon de manière à obtenir des estimations acceptables pour des unités appelées "zones principales de sols" (Major Land Resource Areas - MLRA). Ces zones sont définies en fonction de la nature du sol et du couvert végétal. On en compte environ 180 sur le territoire visé par l'enquête. Par ailleurs, la superficie estimée pour chaque comté doit concorder avec la superficie totale du comté. L'échantillon de l'enquête comprend environ 3,100 comtés. Comme il doit y avoir concordance entre les estimations de superficie relatives aux comtés et celles relatives aux zones principales de sols, l'unité de totalisation fondamentale est la portion d'une zone principale de sols comprise dans les limites d'un comté. Ces unités de base sont au nombre de 5,530 et sont désignées par le sigle MLRAC.

Le plan de sondage équivaut à la forme la plus élémentaire d'une enquête par panel puisque l'échantillon de 1987 est à peu de choses près un sous-ensemble de l'échantillon de 1982. On a choisi d'utiliser comme variables de contrôle les superficies représentées en 1982 par 14 modes d'utilisation du sol parmi les plus importants (par ex.: terrain labourable, terrain de parcs, terrain forestier et terrain urbain). De plus, les données externes, comme la superficie représentée par les routes en 1987, et les données des segments, comme la superficie urbaine en 1987, constituent de l'information supplémentaire au même titre que les données tirées des enregistrements incomplets.

Le tableau 4 est la version condensée d'un tableau d'estimations pour un des États visés par l'enquête. On n'y retrouve que 4 des 14 modes d'utilisation du sol considérés pour l'estimation. Les chiffres figurant dans la colonne de droite sont les estimations pour 1982. Les totaux des colonnes 3 (terrain urbain) et 4 (routes) sont tirés respectivement des données des segments et des sources externes. Le vecteur formé des quatre premiers totaux de la colonne de droite et des deux derniers totaux de la ligne du bas (superficie de terrain urbain en 1987 et superficie représentée par les routes en 1987) est un vecteur de totaux qui correspond au vecteur des moyennes estimées, μ_X , de la section 3.

Les autres chiffres du tableau sont essentiellement des estimations obtenues par les moindres carrés, qui satisfont les six totaux de contrôle. Au cours du processus d'estimation, on a dû parfois recourir à des méthodes d'imputation, par exemple lorsqu'un changement observé dans les données des segments ne se reflétait pas dans les données des points d'observation.

Le plan de l'enquête s'est traduit par des variances élevées pour les estimations directes de la variation de superficie dans le cas des modes d'utilisation du sol relativement moins

L'efficacité de cette méthode dépend de la corrélation entre les variables de contrôle choisies et la variable d'analyse. Si l'une des variables de contrôle est aussi la variable d'analyse, la méthode sera très efficace. La seule raison pour laquelle cette méthode n'est pas parfaitement efficace pour les variables de contrôle est que la méthode des moindres carrés généralisés n'utilise qu'une quantité limitée de données.

Son principal avantage est qu'elle produit une seule série de données de totalisation à partir de laquelle on peut construire des estimateurs ayant la propriété d'additivité pour toutes les périodes visées par l'enquête et tous les tableaux de variation brute. L'inconvénient est que les estimations pour des périodes précises ne sont pas parfaitement efficaces.

La variance de cette méthode peut être calculée de la même manière que celle de la méthode utilisée pour l'échantillonnage double. Soit Y le caractère d'intérêt. Pour plus de simplicité, nous allons supposer un échantillonnage aléatoire simple à chaque fois. Nous définissons comme suit le modèle servant à l'estimation:

$$Y_i = \mu_Y + (X_i - \mu_X)\theta + e_i$$
$$\mu_X = E\{X\},$$
$$e_i \sim \text{Ind}(0, \sigma_e^2).$$

Soit $\hat{\mu}_X$ l'estimateur par les moindres carrés généralisés de μ_X . Alors, l'estimateur de la moyenne de Y s'écrit

$$\hat{\mu}_Y = \bar{y} + (\hat{\mu}_X - \bar{x})\theta,$$

où θ est le vecteur des coefficients de régression que nous avons calculés en faisant la régression de Y_i par rapport à X_i à l'aide de la série d'enregistrements complets et (\bar{y}, \bar{x}) est le vecteur de moyennes pour les éléments observés à tous les passages. Posons m comme le nombre de ces enregistrements. Alors, la variance de l'estimateur est approximativement

$$V\{\hat{\mu}_Y\} = m^{-1}\sigma_e^2 + \theta'V\{\hat{\mu}_X\}\theta,$$

où $V\{\hat{\mu}_X\}$ est la matrice des covariances de $\hat{\mu}_X$.

L'estimateur que nous venons de décrire sera efficace dans la plupart des circonstances. Cependant, il se peut qu'il produise des valeurs estimées négatives pour des quantités réputées non négatives parce qu'il est linéaire et que certains poids peuvent être négatifs. Des méthodes ont été mises au point pour corriger cette lacune. Voir Huang et Fuller (1978).

4. INVENTAIRE DES RESSOURCES NATIONALES DES E.-U.

Le Iowa State Statistical Laboratory collabore avec le U. S. Soil Conservation Service à la réalisation d'une enquête d'envergure sur l'utilisation du sol aux Etats-Unis. Des enquêtes ont déjà eu lieu en 1958, 1967, 1975, 1977, 1982 et 1987. On en prévoit une autre en 1992.

Cette enquête permet de recueillir des données sur la nature et l'utilisation du sol, le couvert végétal, la possibilité de transformer des terres qui ne servent pas actuellement à la culture en terres labourables, l'érosion hydrique et les méthodes de conservation. La collecte des données est confiée à des employés du U.S. Soil Conservation Service tandis que l'université Iowa State s'occupe de l'élaboration du plan de sondage et de l'estimation.

L'échantillon est un échantillon stratifié des terres non fédérales de 49 Etats (l'Alaska étant exclu) et de Porto Rico. Les unités d'échantillonnage sont des portions de terrain appelées segments. La superficie de ces segments varie de 40 à 640 acres. Des données sont recueillies

La seconde méthode consiste à établir des estimations pour chaque période à l'aide des données dont on dispose pour la période en question. Cette méthode est souvent utilisée pour les enquêtes périodiques; les résultats sont publiés à la fin de chaque enquête et ne sont pas révisés par la suite et aucune estimation longitudinale n'est produite. Un des avantages de cette méthode est qu'il est très facile d'établir des estimations pour la période t puisqu'on ne se sert pas des données de la période précédente pour calculer les estimations des valeurs courantes. Avec cette méthode, on obtient habituellement des estimations acceptables (non optimales) des valeurs courantes mais des estimations de la variation qui laissent à désirer.

Par ailleurs, on peut utiliser les deux méthodes dans une même enquête. La Survey of Income and Program Participation (SIPP), réalisée par le U.S. Bureau of the Census, est une enquête par panel avec rotation du moment d'interview et une période de rappel de quatre mois. À chaque passage de l'enquête, le U.S. Bureau of the Census produit une série de poids qui peuvent servir à établir des estimations pour la période en question à l'aide des données fournies par l'ensemble des personnes qui ont participé à l'enquête à cette occasion. L'organisme américain produit aussi a) l'échantillon de personnes qui ont participé aux huit passages de l'enquête en 1984-1985 de même que les poids relatifs à ces personnes, b) l'échantillon de personnes qui ont participé aux quatre passages de l'enquête en 1984 de même que les poids correspondants et c) l'échantillon de personnes qui ont participé aux quatre passages en 1985 de même que les poids correspondants.

Nous allons maintenant décrire une méthode d'estimation pour une enquête par panel où il y a des cas de non-réponse et où l'analyse n'est réalisée qu'à la toute fin. Nous supposons qu'une proportion raisonnable des unités participent à tous les passages de l'enquête et que l'analyse longitudinale revêt de l'intérêt. Il s'agit ici de construire des poids pour les unités dont les enregistrements sont complets. Les données fournies par les répondants dont les enregistrements sont incomplets servent d'information supplémentaire.

La première étape de l'analyse consiste à choisir quelques variables qui ont une grande importance pour l'enquête. Le nombre de variables que l'on peut utiliser dépendra de la taille de l'échantillon. Dans un deuxième temps, on calcule la structure des covariances du vecteur des estimations, qui est composé des estimations simples calculées pour chacune des variables pour chaque genre de schéma de réponse et chaque période pertinente. La structure des covariances est une fonction du schéma de réponse et de non-réponse. Il existe plusieurs définitions des estimateurs simples. Dans le cas de l'échantillonnage aléatoire simple, les estimateurs simples sont des moyennes. Dans le cas de l'échantillonnage stratifié, le vecteur initial peut être défini de manière à inclure les estimations pour chaque strate. Par ailleurs, l'estimateur simple pourrait servir à pondérer les réponses dans chaque strate pour compenser la non-réponse. Le vecteur X figurant dans l'équation (1) est justement un vecteur d'estimations simples.

Étant donné le vecteur d'estimations simples et la matrice des covariances estimée de ce vecteur, nous pouvons construire à l'aide des moindres carrés généralisés de meilleurs estimateurs pour chacune des périodes. Par exemple, si nous avons une enquête par panel avec trois passages, il y a sept schémas de réponse possibles, soit XXX , $XX0$, $XX0$, $XX0$, $XX0$, $XX0$, où X signifie réponse et 0 , non-réponse. Si nous choisissons deux variables d'intérêt, le vecteur des estimations simples contiendra $12 \times 2 = 24$ estimations puisqu'il y a 12 réponses – groupes associés aux sept schémas de réponse. Dans cet exemple, les moindres carrés généralisés serviraient à produire six estimations, soit les valeurs estimées des deux variables d'intérêt pour chacune des trois périodes.

L'estimateur par les moindres carrés généralisés des caractères choisis sert de variable de contrôle à une étape ultérieure. À l'aide de méthodes de régression, nous construisons des poids pour les personnes qui ont participé à tous les passages. Ces poids sont construits de telle manière que les estimations par les moindres carrés généralisés pour chaque période visée soient reproduites par l'échantillon pondéré des personnes qui ont participé aux trois passages. En d'autres termes, les valeurs estimées des variables choisies pour les diverses périodes servent de données de contrôle.

où $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_b)$, $\Gamma' = (\Gamma_1', \Gamma_2', \dots, \Gamma_b')$ et $g' = (g_1, g_2, \dots, g_b)$. Si nous remplaçons g par la combinaison linéaire $G\gamma$, l'équation ci-dessus devient

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \lambda \end{pmatrix} = \begin{pmatrix} X'V^{-1} \\ \Gamma \end{pmatrix} Y.$$

Cette équation définit l'estimateur restreint de β comme une fonction linéaire de Y . Par conséquent, la variance de l'estimateur de β correspondra à la partie supérieure $k \times k$ de

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ \Gamma \end{pmatrix} \begin{pmatrix} G \\ 0 \end{pmatrix}.$$

Cette méthode n'est pas la seule façon de calculer l'estimateur par les moindres carrés généralisés restreint. Il y a aussi l'estimateur composite, qui est un autre estimateur de niveau et de variation qui ne modifie en rien l'estimateur précédent. Voir par exemple Wolter (1979).

Ce court exemple met en lumière plusieurs points. Premièrement, compte tenu d'une corrélation de 0,591 entre les niveaux d'emploi des deux périodes, l'application des moindres carrés généralisés améliore modérément (environ 10 %) l'estimation du niveau d'emploi pour la période courante. En revanche, la même méthode améliore sensiblement la variance de la valeur estimée de P^{EE} . En effet, cette variance équivalait à environ 45 % de la variance de l'estimateur simple. Le second point à retenir est que l'application des moindres carrés généralisés avec contrainte donne des estimations de P^{EE} et P^E qui sont presque aussi efficaces que celles obtenues par les moindres carrés généralisés sans contrainte. Dans le cas de P^E , la perte d'efficacité est d'environ 1 % et dans le cas de P^{EE} , elle est d'environ 6 %.

3. ESTIMATEURS LONGITUDINAUX

Nous avons défini plus haut l'enquête à échantillon constant comme une enquête où les mêmes éléments sont observés à chaque période de collecte des données. L'enquête à échantillon constant se prête bien à l'observation de certaines unités physiques, comme des parcelles de terrain. Par contre, en ce qui a trait à l'observation de populations humaines, l'enquête à échantillon constant n'est rien de plus qu'une vue de l'esprit. Dans la réalité, l'enquêteur perd toujours une partie du groupe de répondants entre deux passages d'une enquête. Lepkowski (1989) et Little et Su (1989) font une bonne analyse des méthodes de traitement de la non-réponse. Voir aussi à ce sujet Little et Rubin (1987), Kalton (1983) et Madow et coll. (1983). Nous avons aussi défini l'enquête avec renouvellement de l'échantillon, où certains éléments de l'échantillon font place à de nouveaux éléments à chaque passage de l'enquête. Dans ce cas, nous pouvons dire qu'il existe une forme de planification de la non-réponse pour les éléments qui sont supprimés de l'échantillon. Il faut donc voir un lien entre l'estimation en situation de non-réponse et l'estimation dans les enquêtes avec renouvellement partiel de l'échantillon. Comme il est difficilement concevable qu'un enquêteur obtienne des réponses de tous les membres de l'échantillon à chaque passage de l'enquête, il faut s'attendre à recourir à une méthode qui permettra de compenser la non-réponse (prévue ou non prévue). Il existe deux méthodes simples et courantes. Si l'intention première de l'enquêteur est de suivre l'évolution d'un groupe de personnes dans le temps, très souvent il ne considérera dans son échantillon que les personnes qui ont participé à tous les passages de l'enquête. Dans ces circonstances, il dispose d'une méthode de pondération par laquelle il peut redresser les données de l'enquête à l'aide des caractéristiques du groupe initial des répondants ou de données supplémentaires ou les deux. On procède souvent de cette façon dans les enquêtes spéciales portant sur une population spécifique. Dans ce cas, les résultats ne sont publiés qu'une fois l'enquête terminée.

de l'échantillon de la première période pour estimer la proportion d'éléments de la classe 1 (personnes avec emploi) à cette période. Pour estimer la proportion d'éléments de la classe 1 aux deux périodes, la méthode ordinaire n'utilise que les éléments communs aux deux échantillons et pour estimer la proportion d'éléments de la classe 1 à la période 2, elle n'utilise que l'échantillon observé à la période 2. Par conséquent, si nous avons un échantillon de 200 éléments à chaque période, l'échantillon de la première période sert à estimer la proportion d'éléments de la classe 1 à cette période, les 100 éléments communs aux deux périodes servent à estimer la proportion d'éléments qui demeurent dans la classe 1 d'une période à l'autre et les 200 éléments observés à la période 2 servent à estimer la proportion d'éléments de la classe 1 pour cette période.

La dernière colonne du tableau 3 contient les variances du meilleur estimateur linéaire sans biais construit à l'aide des moindres carrés généralisés. Cet estimateur est construit à l'aide du vecteur des cinq estimateurs de base et de la matrice des covariances de ce vecteur. Sa formulation est la suivante:

(1)
$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y, \text{ où } V \text{ est définie dans le tableau 2, } \beta = (P_E, P_E, P_{EE}),$$

$$X' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

et V est le vecteur quinquidimensionnel des estimations directes,

$$V' = (\bar{P}_{E.1}, \bar{P}_{E.2}, \bar{P}_{EE}, \bar{P}_{E2}, \bar{P}_{E3}).$$

La seconde colonne du tableau 3 contient les variances de l'estimateur par les moindres carrés restreint, la contrainte étant que l'estimateur pour la période 1 doit être l'estimateur obtenu à l'aide de l'échantillon initial. Cette méthode conviendrait si les organismes statistiques ne révisaient jamais les estimations déjà publiées. Par exemple, le Bureau of Labour Statistics des Etats-Unis ne révise jamais les statistiques du chômage. Une fois publiées, ces statistiques tiennent lieu d'estimations officielles. Il est vrai toutefois que, par rapport à notre exemple, ces statistiques reposent sur un échantillon plus complexe et une enquête qui s'étend sur une plus longue période.

Pour décrire l'estimateur par les moindres carrés généralisés restreint du tableau 3, définissons le modèle

où X est une matrice fixe $n \times k$ et

$$Y = X\beta + e,$$

L'estimateur par les moindres carrés généralisés de β , dont certains éléments sont contraints à être des combinaisons linéaires de X , peut être construit de la façon suivante. Considérons la fonction lagrangienne

$$(Y - X\beta)'V^{-1}(Y - X\beta) - 2 \sum_{i=1}^b \lambda_i(\Gamma_i'\beta - g_i),$$

où Γ_i est un vecteur ligne fixe et b est le nombre de contraintes. La solution à ce problème de minimisation est définie

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} \beta \\ g \end{pmatrix} = \begin{pmatrix} X'V^{-1}Y \\ g \end{pmatrix},$$

variation de niveau d'une période à l'autre. Supposons aussi que nous voulons dresser un tableau des variations brutes; cette opération suppose l'estimation des fréquences par case du tableau de contingence. Pour un tableau 2×2 , il suffit d'estimer la fréquence de la case (1, 1) et les proportions marginales pour obtenir les fréquences des autres cases.

Notre analyse porte sur deux périodes, pour lesquelles le même nombre d'éléments sont observés. Nous supposons que la moitié des éléments observés à la première période le sont aussi à la seconde. Autrement dit, les éléments observés à la seconde période se répartissent en deux groupes égaux: un groupe formé d'éléments observés à la première période et l'autre formé de nouveaux éléments. Le vecteur des observations est composé de la proportion d'éléments de la classe 1 qui font partie de l'échantillon observée uniquement à la première période [désignée par $P_{E.1}$], de la proportion d'éléments de la classe 1 qui font partie de l'autre moitié de l'échantillon de la période 1 [désignée par $P_{E.2}$], de la proportion d'éléments de la classe 1 aux deux périodes, qui font partie de la moitié d'échantillon observée aux deux périodes [désignée par P_{EE}], de la proportion d'éléments de la classe 1 à la période 2 parmi les éléments observés aux deux périodes [désignée par P_{E2}] et de la proportion d'éléments de la classe 1 à la période 2 parmi les éléments observés qu'à la période 2

Nous supposons un échantillonnage aléatoire simple. Comme les statistiques consistent en des proportions d'échantillon, on peut reproduire facilement la matrice des covariances du vecteur de cinq estimateurs. Un multiple de cette matrice est représenté par le tableau 2. Pour obtenir la matrice des covariances pour un échantillon de taille n à chaque période, il suffit de diviser chaque élément de la matrice du tableau 2 par n , puis de multiplier le résultat par deux. Le tableau 3 donne les variances obtenues avec diverses méthodes d'estimation. La première colonne contient les variances obtenues avec la méthode qui n'utilise que les éléments

Tableau 2

Matrice des covariances du vecteur des proportions d'échantillon, deux périodes et échantillons se chevauchant dans une proportion de 50% (Pour un échantillon de taille n , multiplier chaque élément du tableau par 2, puis diviser par n)

$P_{E.1}$	$P_{E.2}$	P_{EE}	P_{E2}	P_{E3}
0.0651	0	0	0	0
0	0.0651	0.0637	0.0819	0
0	0.0637	0.0546	0.0546	0
0	0	0	0	0.0564

Tableau 3

Variances obtenues avec diverses méthodes d'estimation (Pour un échantillon de taille n à chaque période, multiplier chaque élément du tableau par 2, puis diviser par n)

Paramètre	Simple	MCG avec contrainte	MCG
P_E	0.0326	0.0326	0.0294
P_{EE}	0.0819	0.0397	0.0374
P_E	0.0278	0.0258	0.0255
P_{EE}/P_E	0.0290	0.0229	0.0220
$P_E - P_E$	0.0429	0.0367	0.0353

optimale des unités entre les groupes d'échantillons chevauchants et non chevauchants. Patterson (1950) a examiné le cas de T sondages successifs avec plusieurs modes de renouvellement partiel des unités. Le plan d'échantillonnage le plus simple prévoyait le renouvellement d'une proportion déterminée des unités d'échantillonnage à chaque nouveau sondage. En outre, Patterson (1950) avait supposé que, pour une unité i donnée, les écarts $x''_i - x_i$, $i = 1, 2, \dots$, suivaient un processus autorégressif du premier ordre, x''_i étant la valeur de l'unité de population i au temps t , et x_i , la moyenne de la population finie correspondante. Suivant le modèle d'erreur qui en a découlé, il a défini des estimateurs optimaux des valeurs fixes x_i et des écarts $x_i - x'_i - 1$. Il s'est également penché sur l'estimation optimale de x_i suivant des formes généralisées du plan de renouvellement partiel, la détermination de la taille optimale de l'échantillon et l'estimation avec erreurs non autorégressives.

La méthode des moindres carrés a été approfondie par Eckler (1955), Gurney et Daly (1965) et Jones (1980). On en est venu aussi à parler d'estimateurs composites; voir à ce sujet Rao et Graham (1964), Graham (1973) et Wolter (1979). Barteaux, Hasabelnaby et Fuller (1989) décrivent comment le Département de l'Agriculture des E.-U. applique la méthode des moindres carrés dans son enquête sur les activités des exploitations agricoles.

Il semble juste d'affirmer que ces auteurs se sont intéressés surtout à des moyennes ou à des totaux pour des périodes précises. Autrement dit, ils n'ont pas étudié explicitement des paramètres longitudinaux comme la proportion d'individus appartenant à une classe particulière à la période 1 et à la période 2. Nous verrons toutefois que la méthode des moindres carrés s'applique à des paramètres de ce genre.

Une caractéristique intéressante de la méthode des moindres carrés linéaires est que les estimateurs relatifs à un certain nombre de caractères ont la propriété d'additivité, c'est-à-dire que la somme de l'estimateur par les moindres carrés de Y et de l'estimateur par les moindres carrés de Z est égale à l'estimateur par les moindres carrés de $Y + Z$. Toutefois, si l'on se sert d'autres vecteurs d'observations pour construire des estimateurs, la propriété d'additivité disparaît. Dans beaucoup d'enquêtes, on ne peut calculer les estimateurs par moindres carrés optimaux pour toutes les périodes parce qu'on ne peut se servir de toute l'information disponible pour l'estimation. Premièrement, il n'est pas possible d'intégrer toutes les données des enquêtes des périodes antérieures à une analyse par les moindres carrés pour la période courante car le nombre de variables dépassera souvent le nombre d'observations. Deuxièmement, l'organisme qui publie les données peut être tenu de respecter un plafond en ce qui concerne le nombre de fois où il est permis de réviser des estimations antérieures. Smith et Holt (1989) se sont penchés sur ce dernier point.

Afin d'illustrer ces problèmes d'estimation, nous avons voulu utiliser un exemple simple. À cette fin, le tableau 1 représente un tableau de contingence qui montre la division de la même variable en deux classes pour deux périodes données, et dont les observations reposent sur un très grand échantillon. Nous avons identifié les classes de ce tableau en désignant la première comme les personnes avec emploi et la seconde comme les chômeurs. Nous supposons que la population ne varie pas d'une année à l'autre. Si nous devons considérer les naissances et les décès, il nous faudrait alors un tableau 3×3 . Supposons que nous voulons estimer la

Tableau 1
Proportions hypothétiques pour deux périodes données

PÉRIODE 1			
Personnes avec emploi	Personnes avec emploi	Chômeurs	
		Chômeurs	Total
Personnes avec emploi	0.91	0.02	0.93
Chômeurs	0.03	0.04	0.07
Total	0.94	0.06	1.00

- D. mesurer des éléments de la variation, dont
- i) la variation brute
 - ii) la variation pour une unité
 - iii) la variabilité pour une unité
- E. produire des données agrégées sur les unités prises individuellement
- F. déterminer la fréquence, le moment et la durée d'événements
- G. accumuler des données sur des populations peu courantes.

Bien que cela ne soit pas explicite, plusieurs de ces objectifs supposent l'estimation des paramètres de modèles spécialisés.

Par ailleurs, Duncan et Kalton définissent quatre genres d'enquêtes: 1) l'enquête à passages répétées, où rien n'est fait pour veiller à ce que des éléments particuliers de la population fassent partie de l'échantillon plus d'une fois, 2) l'enquête à échantillon constant, où les mêmes éléments sont observés à chaque période, 3) l'enquête avec renouvellement de l'échantillon, où des éléments de la population sont observés pour un nombre déterminé de périodes, puis supprimés, de l'échantillon par renouvellement selon un plan déterminé, et 4) l'enquête à panel fractionné, qui est une combinaison de l'enquête à échantillon constant et de l'enquête 1) ou 3). Duncan et Kalton indiquent aussi sous forme de tableau les genres d'enquêtes qui conviennent le mieux aux différents objectifs.

Lorsqu'un établissement réalise une enquête à passages répétées, il doit parer à toutes les difficultés qui accompagnent normalement l'exécution d'une enquête unique sauf que dans ce cas-ci, les problèmes sont amplifiés. Pour assurer la qualité d'une enquête à passages répétées, il est nécessaire de procéder toujours de la même façon sur le terrain et d'appliquer les mêmes méthodes de traitement, de gestion de données et d'estimation pour toutes les périodes. Il est difficile d'obtenir la collaboration constante des répondants pour plusieurs périodes successives et tout aussi difficile de retrouver les répondants qui ont déménagé. L'erreur de réponse est présente dans tous les genres d'enquête sauf que dans le cas des enquêtes à passages répétées, il faut composer avec un phénomène de "conditionnement" lié à la répétition des interviews. De plus, les erreurs de réponse ont pour effet de créer des incohérences dans les données lorsque celles-ci sont recueillies sur une longue période. Enfin, le changement de composition des unités, telles les familles, vient compliquer à la longue l'estimation et l'analyse.

Nous n'aborderons ici que quelques-unes des questions qui se rattachent aux enquêtes à passages répétées. Notre analyse est fondée sur une grande enquête réalisée par le U.S. Soil Conservation Service en collaboration avec l'université Iowa State. Dans la section 2, nous examinons quelques-unes des méthodes d'estimation utilisées dans les enquêtes à passages répétées. Cette analyse se prolonge dans la section 3, où il est surtout question de l'estimation de paramètres longitudinaux dans les enquêtes par panel. Dans la section 4, nous exposons brièvement les méthodes d'estimation utilisées dans l'enquête du U.S. Soil Conservation Service. Enfin, la section 5 renferme une brève description des effets de l'erreur de mesure sur les estimations de la variation brute.

2. ESTIMATION

Dans cette section, nous allons exposer à grands traits la méthode d'estimation par les moindres carrés généralisés appliquée à des enquêtes où seul un sous-ensemble des éléments de l'échantillon est observé pendant des périodes consécutives. La méthode des moindres carrés généralisés est la première méthode à laquelle se sont intéressés les auteurs qui étudiaient l'estimation dans les enquêtes à passages répétées. Sur les traces de Cochran (1942), Jessen (1942) fut le premier à envisager la construction de poids à variance minimum pour une série d'estimateurs non biaisés établis pour chaque période visée par l'enquête.

Jessen (1942) a analysé le cas particulier de l'échantillonnage effectué à deux reprises où le nombre d'observations diffère d'un échantillon à l'autre et s'est intéressé à la répartition

Analyse d'enquêtes à passages répétés

WAYNE A. FULLER¹

RÉSUMÉ

Dans cet article, nous nous intéressons principalement aux enquêtes à passages répétés où une partie des unités de l'échantillon est observée sur plusieurs périodes et une partie n'est pas observée à certaines périodes. Nous voyons en quoi consiste l'estimation par les moindres carrés pour de telles enquêtes. Nous nous arrêtons aussi à des méthodes d'estimation en vertu desquelles les estimations existantes n'ont pas à être révisées lorsque de nouvelles données sont connues. Par ailleurs, nous considérons des méthodes pour estimer des paramètres longitudinaux; mentionnons à cet égard les tableaux de variation brute. Nous décrivons aussi la méthode d'estimation utilisée dans une enquête à passages répétés sur l'utilisation du sol, réalisée par le U.S. Soil Conservation Service. Enfin, nous illustrons l'effet de l'erreur de mesure sur les estimations de la variation brute et montrons que les plans de sondage qui permettent d'estimer les paramètres du processus d'erreur de mesure peuvent être très efficaces.

MOTS CLÉS: Échantillon d'enquête; moindres carrés; erreur de mesure; variation brute.

1. INTRODUCTION

L'analyse d'enquêtes à passages répétés suscite beaucoup d'intérêt. Soulignons à cet égard la publication récente des actes d'un symposium sur les enquêtes par panel, colligées par Kasprzyk, Duncan, Kalton et Singh (1989), la tenue de séances sur la question lors des assemblées de l'Institut international de Statistique de 1987 et de 1989, et le Symposium sur l'analyse des données dans le temps, organisé par Statistique Canada en octobre 1989. Dans l'article qu'ils ont présenté à la session de l'IIS de 1989 à Paris, Smith et Holt (1989) parlent d'un intérêt renouvelé pour l'élaboration et l'analyse d'études longitudinales. Ils soulignent que des spécialistes de domaines comme la sociologie et la santé réalisent depuis longtemps des enquêtes par panel et des études de cohorte. Ils citent Lazarsfeld et Fiske (1938). Dans le domaine de la santé, mentionnons l'article de Garcia, Battese et Brewer (1975). Les organismes officiels réalisent de nombreuses enquêtes périodiques, comme l'enquête sur la population active. Ces enquêtes produisent habituellement une suite de rapports comme ceux portant sur l'emploi et le chômage pour la période courante. En règle générale, les enquêtes réalisées par les organismes officiels fournissent très peu de données sur le comportement des unités de l'échantillon dans le temps. La U.S. Survey of Income and Program Participation est un exemple d'enquêtes qui servent à produire des estimations longitudinales. Voir à ce sujet Kasprzyk et McMillen (1987). Bien que nous en sachions moins sur les enquêtes réalisées par le secteur privé que sur celles réalisées par les administrations publiques, il semble que les premières servent surtout, comme les secondes, à produire une suite de rapports pour des périodes données. Toutefois, le secteur public comme le secteur privé doivent répondre à une demande accrue d'analyses longitudinales.

L'élaboration d'une taxinomie pour les enquêtes à passages répétés a pour effet de mettre en relief les questions complexes qui accompagnent ce genre d'enquêtes. Duncan et Kalton (1987) énumèrent sept objectifs des enquêtes à passages répétés, soit:

- A. produire des estimations de paramètres de la population pour des périodes déterminées;
- B. produire des estimations de paramètres de la population pour des périodes combinées;
- C. mesurer la variation nette;

¹ Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa, 50011, E.-U.

Une mauvaise datation, ou "télécopage", est une source d'erreurs bien connue dans les enquêtes rétrospectives. Silberstein estime les effets du télécopage pour obtenir des estimations de la première vague non bornée dans l'enquête sur les dépenses des consommateurs (Consumer Expenditure Interview Survey) aux États-Unis. Elle constate que les estimations de la première vague sont plus grandes que celles des vagues suivantes, même après prise en compte des effets de télécopage et conclut qu'une réduction de la période de rappel de la première vague améliore les déclarations dans les vagues suivantes.

Slasny présente plusieurs modèles des flux bruts en présence de non-réponse. Les modèles sont partagés en modèles avec probabilités de transition symétriques et non symétriques. Des méthodes pour l'obtention des estimations des paramètres des différents modèles sont élaborées et appliquées aux données de la victimisation provenant de l'enquête nationale sur la criminalité (National Crime Survey) aux États-Unis.

Enfin, le lecteur remarquera qu'à compter de ce numéro, *Techniques d'enquête* a une nouvelle présentation. L'ancienne couverture était utilisée depuis décembre 1984 (Vol. 10, n° 2). Statistique Canada est en train de modifier de la même façon toutes ses publications afin d'y incorporer un logo unique et d'uniformiser ainsi leur présentation.

Le rédacteur en chef

Dans ce numéro

Ce numéro contient une section spéciale sur les méthodes des séries chronologiques dans les enquêtes, sujet qui a suscité un intérêt considérable ces dernières années. Nous tenons à remercier tout particulièrement W.A. Fuller et J.N.K. Rao pour avoir bien voulu coordonner la rédaction de cette section.

Les deux premières communications de cette section spéciale traitent des problèmes du plan de sondage et de son maintien ainsi que de l'estimation des divers paramètres visés dans les enquêtes répétées. Fuller remarque que des enquêtes répétées destinées à permettre l'estimation des paramètres du processus de mesure des erreurs peuvent se révéler très économiques. Dans le cas d'une enquête de deux périodes avec un chevauchement de 50 %, il montre que des estimations des moindres carrés généralisés des paramètres longitudinaux peuvent avoir une variance sensiblement plus basse que l'estimateur simple basé uniquement sur les unités se chevauchant. Wolter et Harter traitent du problème du maintien d'un échantillon pour une enquête récurrente. L'emploi judicieux de la courbe de Peano permet ainsi d'obtenir plusieurs propriétés souhaitables. Ils présentent une application à une enquête de marché.

Bell et Hillmer examinent la philosophie inhérente de la méthode des séries chronologiques pour l'estimation d'enquêtes répétées compte tenu de la prise en compte de deux sources de variation: la variation des séries chronologiques et la variation de l'échantillonnage. Ils obtiennent des résultats théoriques quant à la cohérence de plan des estimateurs des séries chronologiques et la non-corrélation des séries des signaux et des erreurs d'échantillonnage. Ils font également remarquer que l'emploi des résultats de l'extraction des signaux à partir de l'analyse des séries chronologiques peut améliorer les estimations d'enquête par la réduction de leur erreur quadratique moyenne.

Pour les enquêtes répétées, on peut obtenir de meilleures estimations régionales en combinant l'approche habituelle utilisant l'estimation synthétique et les modèles de séries chronologiques. Pfeffermann et Burck examinent les propriétés statistiques de ces prédicteurs. Ils illustrent la procédure par des données sur les prix des maisons vendues. La communication de Binder et Dick traite des séries chronologiques décrites par des modèles de régression ARMMI avec des erreurs d'enquête suivant un processus ARMM. Il est possible d'appliquer ces modèles à des données provenant d'enquêtes à deux degrés, où les unités de premier degré sont replacées de façon aléatoire, tandis que celles de deuxième degré suivent un plan de panel rotatif. Les auteurs se servent de données provenant de l'Enquête sur la population active pour leur exemple.

Brillinger étudie la relation des naissances et du temps et de la géographie en se servant des données pour les femmes âgées de 25-29 ans en Saskatchewan. Les données agrégées par division de recensement permettent d'obtenir des surfaces lisses. La distribution log-normale de Poisson est également ajustée aux données.

Dans la dernière communication de cette section spéciale, Laniel et Fyfe décrivent le problème de l'étalement des séries infra-annuelles et examinent brièvement quelques solutions présentées dans la littérature. Ils exposent ensuite deux nouvelles méthodes, l'une basée sur un modèle pour les tendances et l'autre, pour les niveaux, et examinent leur validité.

Dans sa communication, Bandyopadhyay prouve que, pour une classe d'estimateurs et de plans de sondage, il est possible de laisser de côté les poids d'échantillonnage lorsqu'on estime un ratio. Il présente ainsi une application à un exemple bien connu pour illustrer le résultat et fait une comparaison en prenant un ratio des estimateurs d'Horvitz-Thompson. Dans le cas des enquêtes répétées avec des panels rotatifs, il est essentiel de connaître les corrélations des panels pour certaines analyses statistiques telles que les études des estimateurs composites. Lee présente la méthodologie de l'estimation des corrélations entre estimations de panels dans l'Enquête sur la population active du Canada.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 16, numéro 2, décembre 1990

TABLE DES MATIÈRES

Dans ce numéro	175
Séries chronologiques dans les enquêtes	
W.A. FULLER	
Analyse d'enquêtes à passages répétés	177
K.M. WOLTER et R.M. HARTER	
Mise à jour des échantillons basée sur les valeurs de Peano	191
W.R. BELL et S.C. HILLMER	
Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques	205
D. PFEFFERMAN et L. BURCK	
Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales	229
D.A. BINDER et J.P. DICK	
Méthode pour l'analyse des modèles ARMMI	251
D.R. BRILLINGER	
Modélisation spatiale et temporelle de données agrégées sur les naissances	267
N. LANIEL et K. FYFE	
Étalonnage des séries économiques	283
S. BANDYOPADHYAY	
Estimation d'un rapport sans connaître le plan d'échantillonnage	291
H. LEE	
Estimation des coefficients de corrélation de panel pour l'Enquête sur la population active du Canada	297
A.R. SILBERSTEIN	
Effets du premier cycle d'interviews dans la Consumer Expenditure Interview Survey aux E.-U.	307
E.A. STASNY	
Symétrie des flux, en tenant compte de la non-réponse, dans les catégories d'actes criminels déclarés par les victimes	321
Remerciements	349

MORRIS H. HANSEN

(1910-1990)

Le présent numéro est dédié à la mémoire de Morris H. Hansen.
Il fut à la fois pionnier, innovateur et meneur
dans le domaine de la méthodologie d'enquête auquel
il a contribué de façon essentielle et durable.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

M.P. Singh

D. Roy

R. Platek

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

B. Afonja, *Nations Unies*

D.R. Bellhouse, *U. of Western Ontario*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

J.-C. Deville, *INSEE*

D. Drew, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of*

Management and Budget

Rédacteurs adjoints

J. Gambino, L. Mach et A. Thèberge, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratiques, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

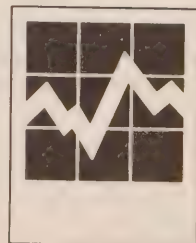
Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (E.-U.) aux États-Unis, et de 49 \$ (E.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

1300220031

Techniques d'enquête

Une revue de Statistique Canada

Décembre 1990 Volume 16 Numéro 2



Statistique Canada
Division des méthodes d'enquêtes sociales

Publication autorisée par le ministre de
l'Industrie, des Sciences et de la Technologie

© Ministre des Approvisionnements
et Services Canada 1991

Tous droits réservés. Il est interdit de reproduire ou de
transmettre le contenu de la présente publication, sous quelque
forme ou par quelque moyen que ce soit, enregistrément sur
support magnétique, reproduction électronique, mécanique,
photographique, ou autre, ou de l'emmagasiner dans un système
de recouvrement, sans l'autorisation écrite préalable du ministre
des Approvisionnements et Services Canada.

Mars 1991

Canada : 35 \$

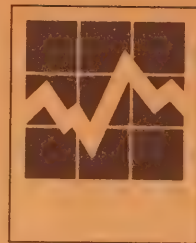
États-Unis : 42 \$ US

Autres pays : 49 \$ US

Catalogue 12-001

ISSN 0714-0045

Ottawa



Catalogue 12-001

Techniques d'enquête

Une revue de Statistique Canada
Décembre 1990 Volume 16 Numéro 2



MAY 20 1992

